THE UNIVERSITY *of* EDINBURGH

The influence of CpG islands on chromatin structure

Elisabeth Wachter

Thesis presented for the degree of Doctor of Philosophy

The University of Edinburgh

2013

Acknowledgments

Firstly, I would like to thank Adrian Bird for all the support he has given me throughout my PhD and also for the desperately needed Friday-night pints after a hard week in the lab. His sense of humour and positive attitude made it definitely easier to go through difficult periods. Also, I would like to thank him for being prepared to swap his dinner with me when, once again, I was unsatisfied with my choice. Special thanks also to Christine for all the administrative help and for making sure I was getting paid.

I am also grateful for all the Bird-lab members, past and present, for making my time in the lab so much fun and for providing so much cake. I would like to thank Thomas who was always there for me and who was always prepared to think about my project and discuss my data. Thanks also to Matt and Martha for providing comments on this thesis and for all their help during my PhD. I would also like to thank Cara for all her help with tissue culture. I am particularly grateful to Sabine, my companion in recombineering-misfortune, for all her help, encouragement and advice and for teaching me how to make perfect chromatin. Without our survival strategy of laughing about failed experiments my PhD would have been definitely more difficult. I would like to thank Jim and Jacky for showing me the tricks of the ES cell targeting-trade. Also, I would like to thank Dina for making sure the lab is running so smoothly. Thanks also to Helene, for all her hugs, wanted or not, and for making the lab a fun place to be in. Thanks to Justyna for her encouragement and for forgiving my tendency to spread to her desk. I would like to thank Timo who made sure I submitted on time by providing a daily countdown. Also thanks to all my fellow PhD students, especially to Sandra, who shared all my up and downs through endless cups of coffees and pints of beers.

I am especially grateful to all of my friends in Edinburgh for many good times kayaking, surfing, kiting and mountain biking, which have helped me regaining energy and motivation for my PhD work. Many thanks also to my family, especially to my parents, for their continuous support. I would like to thank my partner Gregory for all his encouragement patience, and his (pretended) interest in CpG islands. Thanks for a great time in Edinburgh!

Finally, I would like to thank the Wellcome Trust for generous PhD funding.

Declaration

I declare that this thesis was composed by myself, the research presented is my own unless otherwise stated and that this work has not been submitted for any other degree.

Elisabeth Wachter
2013

Abstract

CpG islands (CGIs) are short GC rich sequences with a high frequency of CpGs that are associated with the active chromatin mark H3K4me3. Most occur at gene promoters and are often free of cytosine methylation. Recent work has begun to clarify the functional significance of CGIs with respect to chromatin structure and transcription. In particular, proteins associated with histone-modifying activities, such as Cfp1 and Kdm2a, bind specifically to non-methylated CGIs via their CxxC domains. For example, artificial promoterless CpG-rich sequences integrated at the 3' UTR of genes recruit Cfp1 and generate novel peaks of H3K4me3 in mouse ES cells without apparent RNA polymerase recruitment. There is also evidence that G+C-rich DNA recruits H3K27me3, a gene silencing mark.

In this thesis I am exploring the constraints on DNA sequence and genomic location that are required to impose both H3K4me3 and H3K27me3 at CGI sequences. Showing that the generation of novel peaks of H3K4me3 and H3K27me3 over a promoter-less CpG rich sequence in a gene desert region is independent of it's location in the genome extends earlier findings. These findings suggest that shared features of the primary DNA sequence at CGIs directly influence chromatin modification. Thus CGIs are not passive footprints of other cellular mechanisms, but play an active role in setting up local chromatin structure.

However, the relative contribution of CpG frequency versus G+C content remains unclear. Therefore a sequence was generated that contains low levels of CpGs, comparable to the bulk genome, but has a G+C content similar to that of CGIs (Low CpG / High G+C). When this sequence was inserted into a gene desert neither marks of H3K4me3 or H3K27me3 were formed, indicating the importance of CpGs. Surprisingly, the reverse sequence with a high CpG frequency similar to that of CGIs and a low G+C content similar to that of the bulk genome (High CpG / Low G+C) did not establish H3K4me3 or H3K27me3 either. However, it was found that this sequence becomes heavily methylated in contrast to CGI-like sequences that remained unmethylated when introduced into a gene desert. This finding suggests that a high G+C content is important for keeping CGI-like sequences methylation free. Upon insertion of this High CpG / Low G+C sequence into mouse ES cells that were devoid of the *de-novo* DNA methyltransferases 3a and 3b (Dnmt3a/3b -/-) both H3K4me3 and H3K27me3 marks were established at the inserted sequence. This discovery confirms the importance of CpGs for setting up local chromatin structure.

Lay summary

In mammalian cells DNA needs to be tightly packed in order to fit into the nucleus. The way this is achieved is by winding the double helix around proteins called histones forming a highly ordered structure, called chromatin. Histones can be chemically modified so that the chromatin becomes either less tightly or more tightly packed. This change in chromatin structure is important for essential cell functions. Some genes need to be accessible in a given cell type whereas others need to be silenced, ensuring the right set of genes are expressed at the right time in the right cell type. Chromatin and modified histone proteins play an important role in this process by attracting different proteins, which are responsible for silencing or activating genes. However, it remains unclear how exactly some of these histone modifications are established in the first place.

When looking at the structure of mammalian genes it becomes apparent that many promoters, the start region of genes, contain DNA sequences with a special base composition. These sequences are called CpG islands because they consist of an unusually high frequency of the combination CG, have an overall high content of Gs and Cs and are mostly unmethylated, unlike the rest of the genome which is heavily methylated. It is interesting that right at the start site of genes there are special sequences that aren't found in other regions of the genome. But it is not clear if these CpG islands have an active function by itself or if they are just passive footprints in the genome, remainders of past events. Recent studies in our laboratory suggest that CpG islands are important regulatory structures that influence local chromatin establishment through their special sequence composition. It was shown that special proteins, for example Cfp1, could specifically bind to non-methylated CpGs thereby attracting histone-modifying enzymes.

In this thesis I am testing the influence of CpG island like sequences on chromatin structure by inserting them into gene desert regions, regions in the genome where there are no other genes present. In this way the effect of base composition can be studied in isolation, without having to take into account other influences such as transcription. We found that by introducing an artificial CpG rich and G+C rich sequence into the gene desert the local chromatin structure changed to a more "opened" form indicating that CpG islands can actively influence their chromatin environment. It was not clear what feature of CpG islands, the high CpG frequency or the overall G+C content, is important for their function. Therefore, I modified the sequence by having one construct that had only a few CpGs but as

many G and Cs as in CpG islands. Interestingly, this construct was not able to change chromatin structure supporting the fact that CpG density is the important feature. By changing the sequence to contain as many CpGs as a CpG islands but an overall G+C content like the rest genome I could show that having a high G+C content is important for keeping CpG islands free from DNA methylation, another essential feature of CpG islands. In summary this thesis provides evidence for the importance of CpGs in attracting histone-modifying enzymes and of GC content to protect against DNA methylation.

Table of Contents

Figures

Tables

Abbreviations

5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethyl cytosine
5mC	5-methyl cytosine
AID	Activation-induced cytidine deaminase
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
BAC	Bacterial Artificial Chromosome
BAH	Bromo Adjacent Homology
CAP	CxxC affinity purification
Cfp1	CxxC finger protein 1
CGBP	CpG binding protein
CGI	CpG island
ChIP	Chromatin immunopreciptitation
COMPASS	Complex Proteins Associated with Set1
CpG	CG dinucleotide
CTD	C-terminal domain of RNA polymerase II
DKO	*Dnmt3a/3b* double knock out ES cells
dpc	days postcoitum
DSBH	Double-stranded β-helix
ES cells	Embryonic stem cells
FBS	Foetal bovine serum
Flp	Flippase
Frt	Flippase recognition target
G418	Geneticin
Gsk3	glycogen synthase kinase-3
H3K27me1/2/3	Histone H3 lysine 27 mono-/di-/tri- methylation
H3K36me1/2/3	Histone H3 lysine 36 mono-/di-/tri- methylation
H3K4me1/2/3	Histone H3 lysine 4 mono-/di-/tri- methylation
H3K9me1/2/3	Histone H3 lysine 9 mono-/di-/tri- methylation
HAT	Histone acetyl transferase
HDAC	Histone deacetylases
HMTs	Histone methyltransferases
HOTTIP	HOXA transcript at the distal tip
HP1	Heterochromatin proteins 1
Hprt	Hypoxanthine-guanine phosphoribosyltransferase

IAA	Isoamyl alcohol
IAP	Intracisternal A-particle
ICF	Immunodeficiency, cetromeric instability and facial anomaly
ICR	Imprinting control regions
iPSCs	Induced pluripotent stem cells
JmjC	Jumonji C
Kb	Kilo base
KDM2A	Lysine demethylase 2a
LIF	Leukemia inhibitory factor
lnc RNA	Long non coding RNA
LTR	Long terminal repeats
MBD	Methyl binding domain
MBP	Methyl binding proteins
MDR	Methylation determining regions
Mecp2	Methyl-CpG binding protein
MEF	Mouse embryonic fibroblast
Mek	Mitogen-activated protein kinase kinase
MLL	Mixed-lineage leukemia
MLL1/2	Mixed-lineage leukaemia 1/2
nc RNA	Non coding RNA
NDR	Nucleosome depleted region
NLS	Nuclear localization signal
NURD	Nucleosome remodeling factor
NuRD	Nucleosomal remodeling & deacetylase complex
o/e	Observed over expected
o/n	Over night
ORIs	Origin of replication
PAF1	Polymerase associated factor 1
PcG	Polycomb group
PCL	Polycomb-like
PCNA	Proliferating- cell nuclear antigen
PGC	Primordial germ cell
PHD	Plant homeodomain
PRC1	Polycomb repressor complex 1
PRC2	Polycomb repressor complex 2
PRE	Polycomb response elements
qPCR	Quantitative polymerase chain reaction
RA	Retinoic acid

RFTS	Replication foci targeting sequence
RMCE	Recombination mediated cassette exchange
RNAi	RNA interference
RNA pol II	RNA polymerase II
RT	Room temperature
SET	Su(var)3-9/Enhancer of zeste/trithorax
shRNA	short hairpin RNA
SID	Set1 interaction domain
siRNAs	short interfering RNAs
SUZ12	Su(z)12 Suppressor of Zeste-12
SWI/SNF	Switch/sucrose non-fermentable
TET	Ten-eleven translocation
TRX	Trithorax
TSS	Transcription start site
UV	Ultraviolet
Wt	Wild type
Xi	Inactive X chromosome

1. Introduction

1.1. DNA methylation

DNA can be chemically modified by addition of a methyl-group to the carbon-5 position of the pyrimidine ring of cytosines (see Figure 1.1.1-1). This modification is well conserved in many plant, animal and fungi models and can be maintained through cell division. DNA methylation in mammals occurs preferentially in the context of CpG dinucleotides and is essential for normal development (Li *et al*, 1992; Okano *et al*, 1999). Three conserved enzymes, the DNA methyltransferases (DNMTS) DNMT1, DNMT3A and DNMT3B are responsible for its establishment and maintenance. Functionally, it has been implicated in genomic imprinting, X-chromosome inactivation, maintenance of genome stability, genome defence and transcriptional repression.

1.1.1. DNA methylation across the kingdoms

The prevalence and distribution of DNA methylation in different kingdoms of life is diverse. While vertebrates display a global DNA methylation distribution, invertebrates frequently show a mosaic methylation pattern, characterized by domains of heavily methylated DNA interspersed with stretches devoid of methylation (Suzuki *et al*, 2007; Tweedie *et al*, 1997). Some of the most frequently studied model organisms such as the yeast *Saccharomyces cerevisiae* and the nematode worm *Caenorhabditis elegans* do not have detectable *Dnmt*-like genes and do not show DNA methylation (Gutierrez, 2004; Proffitt *et al*, 1984), whereas DNA methylation is detected at very low levels during early *Drosophila melanogaster* embryogenesis (Lyko *et al*, 2000). Some fungi like *Neurospora crassa* possess DNA methylation, but only repetitive elements are targeted for methylation. Plants are generally heavily methylated but display diverse methylation patterns (Montero *et al*, 1992). While *Arabidopsis thaliana* displays mosaic methylation similar to that of invertebrates with methylated gene bodies and transposable elements, maize, which contains a much larger genome and high levels of transposons, shows a global DNA methylation distribution where genes tend to be unmethylated (Zhang *et al*, 2006).

The transition from mosaic to global CpG methylation occurs at the invertebrate-vertebrate boundary as *Ciona intestinalis*, a sea squirt that is the most vertebrate-like among

invertebrates, displays mosaic DNA methylation pattern (Suzuki *et al*, 2007). In contrast, jawless fish, the most primitive of vertebrates, already show global CpG methylation (Tweedie *et al*, 1997). Whilst invertebrate methylation is mostly restricted to housekeeping genes without preference for repetitive elements, vertebrate methylation occurs at all DNA elements such as genes, indiscriminately of their activity state, transposons and intergenic regions. This ubiquitous DNA methylation makes it difficult to establish if certain DNA sequences are specifically targeted or if all sequences are methylated by default. For example it is unclear if genome defence through specific methylation of transposable elements is a major feature of vertebrate DNA methylation or if these elements are passively methylated as almost the entire genome is methylated (Suzuki & Bird, 2008). It is interesting to speculate that this transition from mosaic to global DNA methylation arose as a consequence of the need to facilitate gene regulation and dampen transcriptional noise in an increasingly complex genome (Bird, 1995).

1.1.1. In mammals DNA is primarily methylated at CpG dinucleotides

As discussed above, the mammalian genome falls into the category of global DNA methylation, which shows a bimodal methylation pattern. The majority, around 70% of all CpGs, are methylated (Ehrlich *et al*, 1982) while a small fraction of the genome (less than 2%) stays unmethylated. These stretches of DNA that co-localize with the majority of promoters are called CpG islands (CGIs) and will be discussed in more detail in section 1.2 (Bird *et al*, 1985).

In mammals, DNA methylation occurs primarily in the context of CpGs, which is a symmetric mark and therefore ensures that both DNA strands are methylated (Ehrlich & Wang, 1981). This provides a mechanism how DNA methylation can be maintained throughout cell division, making it an epigenetic heritable mark that does not alter the coding potential of its underlying DNA sequence. However, a consequence of DNA methylation at CpGs is its mutagenic character. Both, cytosine and 5-methylcytosine are prone to hydrolytic deamination, resulting in the presence of uracil and thymine, respectively (Figure 1.1.1-1). As uracil is a base not found in DNA it is recognized efficiently by DNA glycosylases and reverted to cytosine. In contrast, thymine is a normal DNA base that after deamination is incorrectly base-paired. This mismatch is only inefficiently repaired by the cell's repair system, for example by the glycosylase MBD4 (Hendrich & Bird, 1998). As a result of this mutability CpGs have been lost from the genome over the course of evolution and the observed over expected ratio of CpGs is only 0.21 (Bird, 1980). DNA methylation in other

contexts than CpG have been observed, mainly in embryonic stem cells (ES cells). Non-CpG methylation predominantly occurs at CpA dinucleotides but has also been found at CpTs (Ramsahoye *et al*, 2000). As this type of methylation lacks the symmetry needed for a replication-based maintenance it might have to be continuously established by *de-novo* methyltransferases. Methylation in non-CG contexts show enrichment in gene bodies and depletion in protein binding sites and enhancers (Lister *et al*, 2009).

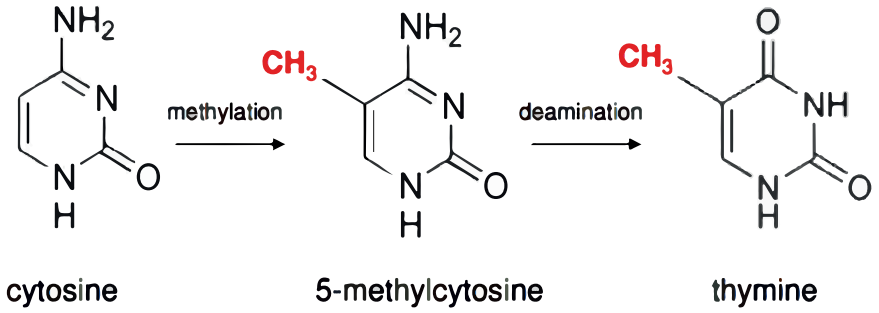


Figure 1.1.1-1 Cytosine methylation
Cytosine can be methylated at the carbon-5 position to form 5-methylcytosine, which can deaminate to thymine.

1.1.2. DNA methyltransferases methylate CpGs

Specific enzymes, so called DNA methyltransferases (DNMTs) are responsible for methylating DNA. All DNMTs use S-adenosyl-methionine (SAM) as the source of the methyl group to be transferred to the DNA bases. Mammalian DNMTs use two different mechanisms to establish DNA methylation with respect to template specificity. One relies on the *de novo* addition of methyl groups to previously unmethylated DNA. The other mechanism re-establishes the DNA methylation pattern after DNA replication to that of the original strand (reviewed in (Hermann *et al*, 2004)). DNMT1, the sole maintenance methyltransferase belongs to the latter class, whereas DNMT3A and DNMT3B are *de novo* methyltransferases. All three DNMTs possess a conserved catalytic domain comprised of 10 highly conserved peptide motives located at the carboxy terminal domain (Figure 1.1.2-1). The observed functional differences between the DNMTs are due to the variable amino-domain containing different motifs.

DNMT1

DNMT1 is a large modular protein composed of a replication foci-targeting sequence (RFTS), a ZF-CxxC domain, a pair of bromo-adjacent homology (BAH) domains and a C-terminal catalytic domain. DNMT1 associates with proliferating-cell nuclear antigen (PCNA) at replication forks via its RFTS, ensuring that the methylation pattern of the parental strand is faithfully transferred to the newly replicated strand after DNA replication (Iida *et al*, 2002; Chuang *et al*, 1997). It associates with the replication machinery and preferentially methylates hemimethylated DNA. This preference for hemimethylated DNA lies in the N-terminal domain as cleavage of this domain abolishes this preference and increases *de novo* methyltransferase activity (Bestor, 1992; Yoder *et al*, 1997). The CxxC domain is probably required to restrict proper targeting of DNMT1 (see also Chapter 1.2.5). Consistent with its function to maintain DNA methylation through replication, it is expressed at high levels in all replicating somatic cells. DNMT1 deficiency in mouse is embryonic lethal, shortly after gastrulation, and is coincident with a 70% reduction in DNA methylation levels. DNMT1-deficient ES cells however can be maintained, leading to the notion that DNA methylation is crucial for development (Li *et al*, 1992). As it was shown that *Dnmt1* knock-out (KO) ES cells are still capable of *de novo* methylating retroviral DNA (Lei *et al*, 1996) it became apparent that there must be a yet unidentified enzyme with DNA methylation ability. Initially it was thought that DNMT2 could fulfil that role but although it possesses all the conserved motives shared by DNMTs, inactivation of DNMT2 does not perturb *de novo* or maintenance methylation (Okano *et al*, 1998a).

DNMT3A/3B

It turned out that DNMT3A and 3B are the long sought after *de novo* methyltransferases that are responsible for establishing DNA methylation patterns *de novo* after global demethylation during embryogenesis and gametogenesis. This is reflected by their expression pattern. Both enzymes are highly expressed in undifferentiated ES cells, are downregulated during differentiation and expressed at low levels in adult somatic cells (Okano *et al*, 1998b). The PWWP domain, a conserved Prolin-Tryptophan-Tryptophan-Prolin motif, of DNMT3A/3B specifically binds histone H3 with trimethylated lysine 36 (H3K36me3), which increases the ability of DNMT3A to methylate nucleosomal DNA (Dhayalan *et al*, 2010). The ADD domain contains many cysteine residues providing interaction with other proteins such as transcription factors, histone deacetylases (HDACs), heterochromatin protein HP1 or histone methyltransferases such as SUV39H1, SETDB1 and EZH2 (Jurkowska *et al*, 2011). Additionally, the ADD has been shown to interact

specifically with the N-terminal part of histone H3 when its lysine 4 residue is not modified (Ooi *et al*, 2007; Zhang *et al*, 2010b). Sequence analysis and domain structure of DNMT3A and DNMT3B suggest overlapping function and indeed double knock-out mice have a more severe phenotype than either single knock-out, suggesting a synergistic effect. Inactivation of both genes blocks *de novo* methylation in ES cells and early embryos, but it has no effect on the maintenance of imprinted DNA methylation patterns. DNMT3A and 3B are essential for normal development as mice deficient for both enzymes die at approximately embryonic day 11.5 (E11.5), however ES cell lines can be successfully maintained (Okano *et al*, 1999). Despite their overlapping function, they also exhibit specific functionality. DNMT3B in particular is known to be required for *de novo* methylation of specific genomic regions, as mice or human patients with DNMT3B mutations are deficient in methylation of pericentromeric repetitive DNA sequences and at CGIs on the inactive X chromosome (Hansen *et al*, 2000). DNMT3A is required for methylation of the major satellite repeats and for establishing maternal imprints through interaction with the third member of the *Dnmt3* family, DNMT3L (Kaneda *et al*, 2004).

DnmtL shows clear homology to *Dnmt3a* and *3b* but lacks essential motives rendering it catalytically inactive. DNTM3A and DNMT3B interact with DNMT3L and this interaction enhances their enzymatic activity (Gowher *et al*, 2005). DNMT3L specifically interacts with the N-terminus of histone H3 and this interaction is strongly inhibited by methylation at lysine 4 of histone H3 but is insensitive to modifications at other positions (Ooi *et al*, 2007). This suggests that DNMT3L might be important in inducing *de novo* methylation by recruiting or activating DNMT3A and 3B.

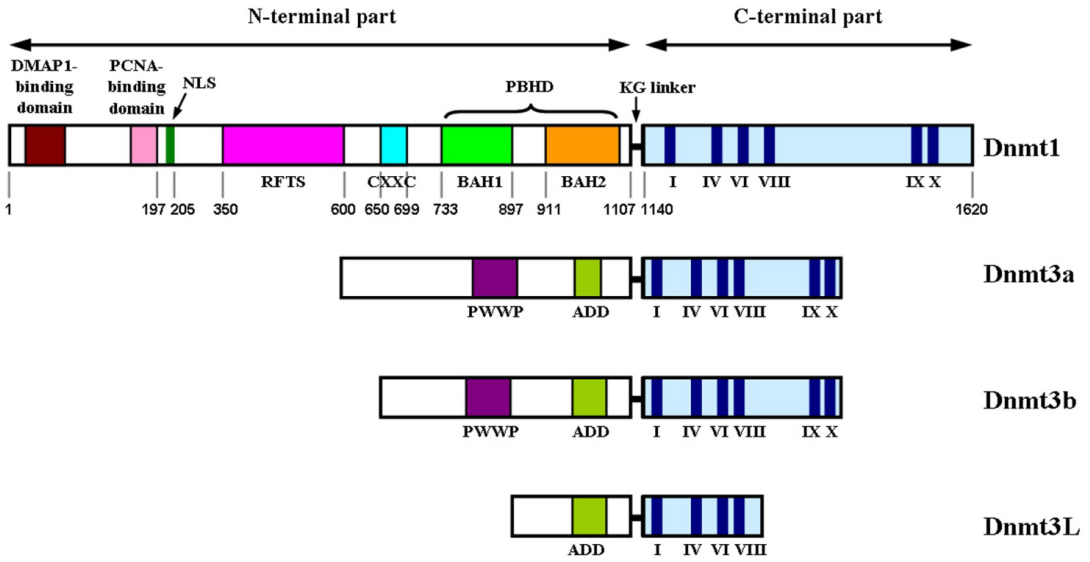


Figure 1.1.2-1 Structure of mammalian DNMTs
The carboxy terminal domain contains the conserved catalytic motifs (roman numerals; dark blue bars) whilst the variable amino terminal region contains various domains required for intracellular delivery and allosteric regulation of the catalytic C-terminus. Adapted from (Ryazanova, 2012).

1.1.3. Dynamics of DNA demethylation

Global changes in DNA methylation occur during normal development, especially during primordial germ cell (PGC) specification and just after fertilization, but also during cell differentiation. During gametogenesis global DNA demethylation occurs over the course of approximately 1 day from E10.5 to E11.5, encompassing the whole genome with the exception of IAP elements and a few LTRs that are only partially demethylated (Seisenberger *et al*, 2012; Popp *et al*, 2010). In the paternal genome methylated cytosine (5mC) is converted to hydroxymethylcytosine (5hmC) at fertilisation, followed by passive loss that reaches a minimum at blastocyst stage (Smith *et al*, 2012) and reviewed in (Smith & Meissner, 2013). Additionally, many tissue specific genes are methylated in almost all tissues but unmethylated in the tissue these genes are expressed in (Illingworth *et al*, 2008; Schilling & Rehli, 2007). From these examples it becomes clear that a demethylation mechanism must exist that either demethylates 5mC globally or specifically at target genes.

For many decades the mechanism of demethylation was unknown and much disputed. Recent work has shown that a direct conversion of 5mC to unmethylated cytosine (C) is unlikely but that this process involves intermediate states, DNA repair and base excision (Guo *et al*, 2011b). 5mC can be modified by the ten-eleven translocation (TET) family of dioxygenases that catalyse the oxidation of 5mC to 5hmC (Ito *et al*, 2010; Tahiliani *et al*,

2009; Koh *et al*, 2011). It has been suggested that deamination may be involved in the process of converting 5hmc to unmodified C, and this is supported by evidence showing that Aid or APOBEC may be necessary for demethylation in the brain (Guo *et al*, 2011a). It is also possible that 5hmC could be deaminated to 5hmU, a good substrate for several known glycosylases (Rusmintratip & Sowers, 2000) but it is still not clear whether 5hmC can actually serve as a substrate for deamination *in vitro* or *in vivo* (Nabel *et al*, 2012). Alternatively, 5hmC can be further oxidized to 5-formylcytosine (5fC) or 5-carboxylcytosine (5caC) (Ito *et al*, 2010; He *et al*, 2011) which can be removed by glycosylation enzymes. Otherwise, 5hmC, which is not recognized by DNMT1, could become diluted out during replication. Passive DNA demethylation might be especially relevant in rapidly dividing cells such as ES cells. The discovery that 5hmC levels are particularly high in brain cells that do not divide supports this model (Kriaucionis & Heintz, 2009). It is also conceivable that 5hmC might function by altering the local chromatin environment through recruitment or displacement of proteins as most 5mC binding proteins do not recognize 5hmC (Valinluck & Sowers, 2007; Jin *et al*, 2010).

The TET protein family consist of TET1, TET2 and TET3, with TET1 being highly expressed in ES cells whereas TET2 and TET3 are more ubiquitously expressed (see also Section 1.2.5). TET1 is a candidate for mediating DNA demethylation during PGC development as it is expressed in PGCs and enriched at germ line specific genes and imprinting control regions which are demethylated at this stage. Global 5hmC levels in PGCs seem to be dependent on TET1 activity (Hackett *et al*, 2013). However, TET1 deficient mice are viable and loss of TET1 does not seem to inhibit germ line competence or embryonic development (Dawlaty *et al*, 2011). Recently, double TET1 and TET2 KO mice were generated and while some embryos show perinatal lethality, also viable mice can be obtained suggesting a compensatory role of TET3 (Dawlaty *et al*, 2013). Adult mutant males have normal gonads and are fertile but double-mutant females display smaller ovaries, reduced numbers of mature follicles, and reduced fertility, which is more pronounced than previously observed in TET1 single-knockout mice. TET3 is expressed at high levels in oocytes and zygotes but is downregulated rapidly from the 2-cell stage onwards. Furthermore, it is enriched in the paternal pronucleus in zygotes and might be the responsible enzyme for demethylation after fertilization, as downregulation of TET3 in the zygote impedes 5mC to 5hmC conversion in the paternal pronucleus (Wossidlo *et al*, 2011; Gu *et al*, 2011). In summary, the recent advances made in deciphering demethylation of DNA highlight the dynamic nature of epigenetic modifications of DNA.

1.1.4. Targeting *de novo* DNA methylation

After erasure of DNA methylation in the early embryo a new pattern of DNA methylation needs to be established. How exactly *de novo* DNMTs are targeted to ensure global methylation with the exception of CGIs has been studied intensively. More and more evidence is emerging that the basic DNA methylation profile during early development is established through interaction with histones. The pattern of H3K4 methylation is probably established before the global wave of *de novo* methylation, possibly through sequence-directed binding of RNA polymerase II (RNA pol II), which then recruits specific H3K4 methyltransferases (Guenther *et al*, 2007). At CGIs, H3K4me3 is established while the rest of the genome is densely packed into nucleosomes that contain unmethylated H3K4. DNMT3L recruits DNMTs to DNA by binding to unmethylated histones. However, interaction of DNMT3L with histones is inhibited by H3K4 methylation. This would provide a mechanism through which the majority of CpGs become methylated while CGIs remain methylation free (Ooi *et al*, 2007; Jia *et al*, 2007).

Once the basic bimodal pattern of DNA methylation has been created, specifically targeted methylation and demethylation events occur throughout development. How this precise targeting occurs is less clear as there is little evidence for sequence specificity for any of the main DNMTs except the central CpG site. If it is not sequence specificity what kind of mechanism does target DNMTs? One proposition is that they are targeted by histone modifying enzymes, especially by histone methyltransferases (HMTs) that are recruited by local regulatory factors (reviewed in (Cedar & Bergman, 2009)). An example for locus-specific gain of DNA methylation is represented by the promoter region of *Oct3/4* (also known as *Pou5f1*), a pluripotency factor, which needs to be shut down for successful development. In ES cell models, it has been shown that this inactivation occurs in a stepwise manner. Initially, transcription is turned off through interaction of repressor molecules with the *Oct3/4* promoter. Subsequently, a complex that contains the HMT G9A and enzymes with histone deacetylase (HDAC) activity are recruited. This leads to deacetylation, which then allows H3K9 methylation (Feldman *et al*, 2006). Finally, the G9A containing complex recruits DNMT3A and DNMT3B to the *Oct3/4* promoter, resulting in *de novo* DNA methylation. Recently it has been shown that the same can be achieved ectopically by artificial targeting of HP1, which instructs H3K9 methylation followed by DNA methylation (Hathaway *et al*, 2012). There is evidence that this mechanism plays a role in the inactivation of other genes such as *Nanog* (Epsztejn-Litman *et al*, 2008).

Targeting of DNMTs to pericentromeric regions is achieved via the H3K9 HMTs SUV39H1/H2. The establishment of H3K9me3 recruits DNMT3B which is responsible for DNA methylation of pericentromeric repeats (Lehnertz *et al*, 2003). Methylation of these repeats is essential as pericentromeric elements contain the potential to initiate transcription. The fact that patients suffering from the heritable, autosomal recessive immunodeficiency, centromeric instability and facial anomaly (ICF) syndrome, display a missense mutation in DNMT3B suggests how the silencing of pericentromeric repeats is essential for normal development (Xu *et al*, 1999).

Alternatively, DNMTs can be recruited to target genes via Polycomb Repressive Complex 2 (PRC2), which is a member of the other major silencing system in cells, the Polycomb system (see section 1.3.3). It has been reported that DNMTs are able to interact with EZH2, the HMT in the PRC2 complex that establishes H3K27me3 (Viré *et al*, 2006). In this study it was suggested that EZH2 is required for DNA methylation of EZH2-target promoters, indicating that EZH2 could act as a recruiting platform. However, it must be noted that usually PRC2 target genes have unmethylated CGI promoters and therefore remain unmethylated throughout development, as H3K27me and DNA methylation seem to be mutually exclusive (Bartke *et al*, 2010). Nonetheless, in certain cases some of these genes become targets of *de novo* methylation, for example during differentiation of ES cells to neuronal precursors (Mohn *et al*, 2008; Meissner *et al*, 2008).

An alternative mechanism of recruiting DNMTs to their targets could involve non-coding RNAs. However, it is likely that targeting of DNMTs via an RNA mediated pathway is not happening directly but again via histone methylating enzymes (Zhao *et al*, 2008; Nagano *et al*, 2008).

The above examples illustrated how histone-modifying enzymes might be important for establishing DNA methylation patterns. Inversely, there is evidence that DNA methylation is needed in order to help to maintain histone modification profiles through cell division. It is probable that histone modification patterns are disrupted during replication and need to be re-established in the daughter cells. As DNA methylation profiles are retained through the activity of DNMT1 on hemimethylated DNA, it might offer a platform used for reconstructing the epigenetic state of the genome following cell division. It is possible that the presence of DNA methylation recruits G9A via its interaction with DNMT1 (Estève *et*

al, 2006). Additionally, it has been shown that DNA methylation inhibits H3K4me3 (Lande-Diner *et al*, 2007).

Another way how DNA methylation leads to a repressive chromatin state involves the recruitment of repressive complexes by methyl binding proteins (MBPs). This family of proteins possesses a methyl binding domain that specifically recognizes and binds to methylated CpGs (Hendrich & Bird, 1998). There are several models, how MBPs repress transcription. The interaction of MBPs with methylated CpGs could either result in steric hindrance of binding of transcription factors or other proteins of the transcriptional machinery. It could occlude transcription factor binding sites or it could recruit histone-modifying enzymes that establish a heterochromatic environment. The MBD proteins consist of MeCP2, MBD1, MBD2, MBD3 and MBD4. For instance, MeCP2 interacts with the SIN3A/HDAC and NCoR/SMRT complexes. MeCP2 is particularly abundant in the brain, specifically in neurons, and mutations in MeCP2 are responsible for the neurological disorder Rett syndrome (Amir *et al*, 1999). Interestingly, some Rett syndrome-causing mutations in MeCP2 abolish the interaction with the NCoR/SMRT co-repressor complexes, which contains HDAC3 (Lyst *et al*, 2013). MBD1 associates with the H3K9 HMTs SETDB1 and SUV39H1 and the heterochromatin binding protein HP1 (Fujita *et al*, 2003), while MBD2 associates with the NuRD co-repressor complex (Zhang *et al*, 1999). MBD3 does not posses a methyl-specific binding domain although it acts as a transcriptional repressor (Saito, 2002). MBD4, which contains a thymine glycosylase domain, is involved in DNA repair. It preferentially repairs T:G mismatches (Hendrich *et al*, 1999). In addition, an unrelated family of *Kaiso*-like proteins also have binding specificity for methylated DNA (Filion *et al*, 2006). Given the severity of symptoms displayed by DNMT knock-out (KO) mice it might be surprising to find that none of the MBD proteins is essential as shown by the relatively mild phenotypes of MBD KO mice, which could suggest functional redundancy between MBD proteins. Alternatively, this might indicate that DNA methylation exerts its effects not only via MBD proteins but also via alternative pathways (reviewed in (Sasai & Defossez, 2009)).

These examples support the notion of bidirectional cross talk of histone modifications and DNA methylation to ensure the establishment and maintenance of a transcriptionally inert chromatin environment.

1.1.5. Role of DNA methylation

In order to understand the function of DNA methylation it is important to realize that its role varies with context and that its repressive influence on transcription might be more complicated and nuanced than initially thought.

Promoter DNA methylation

How promoter DNA methylation leads to transcriptional repression has been extensively studied. Early studies describe how expression of genes is extinguished upon artificially methylation of their promoters (Stein *et al*, 1982; Vardimon *et al*, 1982). Later, genome-wide studies confirmed these results by correlating methylated promoters with transcriptional inactivity (Mohn *et al*, 2008; Weber *et al*, 2007). However, these examples do not show that DNA methylation is the initial silencing signal. Indeed, it is probably true that DNA methylation maintains an already silenced state rather than initiating it. One striking example is the establishment of X-inactivation where DNA methylation occurs only after transcriptional silencing is completed (Lock *et al*, 1987). Another example where silencing precedes DNA methylation is the mechanism by which *Oct3/4* is shut down during differentiation where first transcription is shut down and a heterochromatic environment created and only then DNA becomes methylated (Epsztejn-Litman *et al*, 2008). Genome-wide studies in cancer cells have shown that CGI promoters that are already marked by Polycomb are more likely to become DNA-methylated (Schlesinger *et al*, 2006).

The majority of mammalian genes possess CGI-containing promoters that remain widely unmethylated. However, in situations that are linked with long-term silencing such as X-inactivation and genomic imprinting, certain promoter CGIs do become methylated (discussed in section 1.2). It is not clear how DNA methylation affects non-CGI promoters despite the fact that this category comprises around 45% of all promoters and contains some well-studied genes such as *Nanog* and *Oct3/4*. A number of non-CGI promoters displays tissue-specific methylation patterns, suggesting that DNA methylation might play a role in the establishment and maintenance of tissue-specific expression profiles. Some studies have shown an inverse correlation between DNA methylation and gene expression, as it is the case for CGI promoters (Eckhardt *et al*, 2006; Han *et al*, 2011). Others have reported that CpG-poor promoters can still be expressed when they are methylated (Weber *et al*, 2007). It has been proposed that promoter strength could play a role in allowing expression from sparsely methylated genes (Boyes & Bird, 1992). Strong promoters are thought to be able to disrupt MECP2 binding to methylated, CpG poor promoters, which prevents transcription (Boyes & Bird, 1992).

Gene body methylation

Although historically most focus was given to promoter methylation, the majority of CpGs lies outside of promoters. From early studies it has become clear that gene body methylation occurs at active genes and is not associated with repression (Zhang *et al*, 2006; Suzuki *et al*, 2007). A study that examined the methylation pattern between the active X-chromosome and the inactive one found a positive correlation between gene body methylation and transcription on the active X-chromosome (Hellman & Chess, 2007). A number of reports in human have shown a more general correlation between DNA methylation in gene bodies and gene expression, with highly expressed genes tending to have more intragenic methylation (Ball *et al*, 2009; Rauch *et al*, 2009). However, other studies suggest that gene body methylation does inhibit transcriptional elongation (Lorincz *et al*, 2004).

It has been suggested that the function of gene body methylation is to repress spurious transcriptional initiation outside of promoter regions. Furthermore, many genes contain several transcriptional start sites and it is conceivable that DNA methylation could regulate alternative promoter usage (Maunakea *et al*, 2010). The finding that there is more DNA methylation in exons than in introns and that the change in DNA methylation density occurs at the exon-intron border led to the suggestion that gene body methylation could play a role in splicing (Laurent *et al*, 2010). Fitting to this theory is the fact that exons seem to have more nucleosomes that are preferential sites for DNA methylation, than introns (Chodavarapu *et al*, 2010; Schwartz *et al*, 2009). Also some factors influencing the rate of transcription through RNA pol II pausing, an important feature of the kinetics of splicing, are regulated by DNA methylation (Shukla *et al*, 2011).

DNA methylation in intergenic regions

It has been suggested that DNA methylation in the bulk genome, which mostly consists of repetitive elements derived from transposable elements, plays a role in genome defence to prevent unconstrained transposition (Yoder *et al*, 1997). Indeed, DNMT-deficient mice show increased expression levels of IAP elements (Walsh *et al*, 1998) and human LINE and SINE elements are also de-repressed when DNMTs are lacking (Woodcock *et al*, 1997). However, so far no elevated transposition activity could be detected in DNA methylation depleted cells (Wilson *et al*, 2007), raising the possibility that DNA methylation serves to suppress spurious transcription from cryptic promoter elements (Bird, 1995).

DNA methylation is also implicated in maintaining genome integrity. As discussed above, mutation in DNMT3B can lead to ICF, a syndrome that is characterized by chromosomal segregation defects which are associated with hypomethylated pericentric satellite repeat sequences (Ehrlich, 2003). It is well documented that cancer cells show abnormal DNA methylation patterns, with many non-CGI regions being hypomethylated while some CGI containing promoters become aberrantly methylated, which is associated with genome instability and chromosomal abnormalities (Jones & Baylin, 2007).

1.2. CpG islands

Unlike the bulk genome, which comprises around 99% of the genome and is heavily methylated, CpG islands (CGIs) cover less than 2% of the genome and are usually not methylated. CGIs are characterised by an elevated G+C content and a high density of unmethylated CpGs. They tend to localize to gene promoters and associate with permissive chromatin.

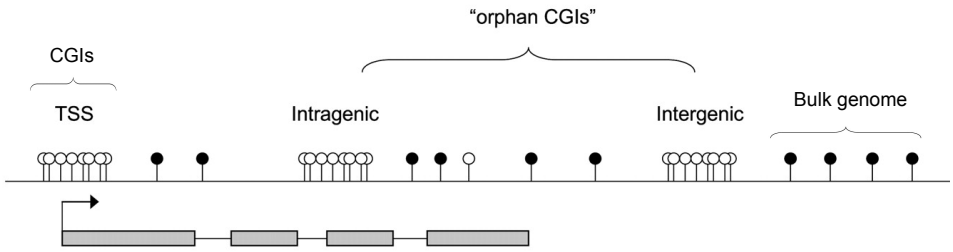


Figure 1.1.5-1 The genomic distribution of CGIs
CGIs can be located at annotated TSSs, within gene bodies (intragenic), or between annotated genes (intergenic). Intragenic and intergenic CGIs of unknown function are classed as "orphan" CGIs. Empty circles: Unmethylated CpG residues. Filled circles: Methylated CpG residues. Adapted from (Deaton & Bird, 2011).

1.2.1. Identification and characterization of CpG islands

CGIs were discovered in the 1980s and initially termed *HpaII* Tiny Fragments (HTFs) due to their characteristic of being cut by the methylation-sensitive restriction enzyme *HpaII*, which recognizes a CCGG sequence (Cooper *et al*, 1983; Bird *et al*, 1985). The majority of mouse genomic liver DNA was resistant to being cut by *HpaII* whereas a small portion, around 1%, was digested by *HpaII* indicating that these sequences were unmethylated. It was suggested that the high content of unmethylated CpGs is a combination of an unusually high G+C content and lack of CpG deficiency due to absence of the mutagenic potential of methylated CpGs. The number of these CpG-rich sequences was initially estimated to be around 30 000 and it was hypothesized that they might be associated with genes (Bird *et al*, 1985).

Originally, these sequences were defined as at least 200bp in length and with a G+C content of 50% and an observed/expected (o/e) CpG frequency of 0.6 (Larsen *et al*, 1992a; Gardiner-Garden & Frommer, 1987). The values suggested by Gardiner-Garden and Frommer are routinely used by genome browsers to predict CGIs. However, these values are somewhat arbitrary and as they are rather relaxed, many false positives are obtained due to the inclusion of repetitive elements. One report suggested the use of more stringent criteria in order to minimize contaminating *Alu* sequences, which have a similar base composition as CGIs and occur in thousands of copies (Takai & Jones, 2002). This increased stringency reduced the number of erroneously included *Alu* elements but also decreased the number of genuine CGIs. Preliminary computational analysis of the human genome sequence identified around 50 000 CGIs (Lander *et al*, 2001), of which only 28 890 were unique when removing multi-copy sequences by screening against known classes of repeats using repeat masking databases. Varying the stringency of parameters greatly changes the number of identified CGIs and any given parameters will always contain an arbitrary element. In order to overcome these limitations in determining the correct number of CGIs, a method was developed in our laboratory that relies on the identification of unmethylated CGIs using a cysteine-rich CxxC3 domain from mouse MBD1 that has a high affinity for nonmethylated CpG sites. This technique is called CxxC affinity purification (CAP) and allows the biological identification of CGIs (Illingworth *et al*, 2008). Using this approach around 17000 CGIs were identified in human blood samples. Using CAP, a comprehensive set of mouse and human CGIs was established in a variety of tissues. Contrary to previous estimates that suggested one third less CGIs in mouse than in human (Mouse Genome Sequencing Consortium *et al*, 2002), the CGI number in mouse and human were found to be very similar, with 25 495 CGIs in humans and 23 021 in mice (Illingworth *et al*, 2010). The reason for this initial discrepancy was found to be the slightly lower overall CpG content of mouse CGIs in comparison to human CGIs. This finding strengthened the notion that CGIs are functionally important sequences as the apparent lack of conservation between mouse and human CGIs initially called their regulatory importance into question.

Traditionally, CGIs were characterized as being associated with gene promoters (Lander *et al*, 2001; Bird *et al*, 1985; Saxonov *et al*, 2006; Weber *et al*, 2007). Approximately 60% of human genes have a CGI promoter including all housekeeping genes as well as approximately 40% of genes with tissue-specific expression patterns (Larsen *et al*, 1992b). CGI promoters often lack TATA boxes and can initiate transcription from multiple positions within the island (Zhu *et al*, 2008). On a genome-wide scale this correlation between

transcription initiation over a few hundred base pairs and presence of CGIs has been confirmed (Carninci *et al*, 2006). Interestingly, half of the CGIs identified in both human and mouse samples do not localize to annotated transcription start sites (TSSs) but are found within genes (intragenic CGIs) and in between genes (intergenic CGIs) (see Figure 1.1.5-1) (Illingworth *et al*, 2008; 2010). It was proposed that all CGIs, independent of their position, act as promoters at some stage in development, as many of these so called "orphan CGIs" show evidence for promoter function (Maunakea *et al*, 2010; Illingworth *et al*, 2010; Macleod *et al*, 1998). Because only a few tissues have been investigated so far for the presence of promoter activity of orphan CGIs and many orphan CGIs are active in a very tissue-specific manner, it is possible that all orphan CGIs are associated with novel transcripts. Some orphan CGIs might be promoters of hitherto uncharacterized protein-coding genes, others may act as alternative promoters of nearby protein-coding genes, while yet others might give rise to noncoding RNAs. Not only the number of CGIs is conserved between mouse and human but also the relative distribution of CGIs to TSSs and inter- and intragenic loci proved to be similar between the two species. The genomic position of many orphan CGIs appears to have been maintained since the divergence of humans and mice, further implying functional importance (Illingworth *et al*, 2010).

1.2.2. Origin and maintenance of CpG islands

As the majority of CGIs stays unmethylated throughout development, the question how CGIs maintain this state of hypomethylation during the global wave of *de novo* methylation is an intriguing one. An abundance of recent findings contributed to our understanding of how this might be achieved. One initial suggestion of how CGIs are kept methylation-free and therefore retain their high density of CpGs was that CGIs might be intrinsically refractory to DNMT-binding due to their DNA sequence. However, given the fact that CpGs seem to be the preferred substrate of DNMTs and that in some cases CGIs do become methylated, i.e. during X-inactivation, this seems unlikely.

Other studies argued that interaction of different protein factors with CGIs might exclude DNMTs and prevent them from targeting CGIs for *de novo* methylation. Evidence for such a mechanism is provided by the fact that mutating a binding site for the ubiquitous transcription factor SP1 in the CGI of the hamster and mouse *Aprt* gene lead to *de novo* methylation (Brandeis *et al*, 1994; Macleod *et al*, 1994). Generally, CGIs show a higher enrichment in transcription factor binding sites compared to the rest of the genome due to their elevated C+G content (Stadler *et al*, 2011).

Gaining more insight into the characteristics of CGIs it became apparent that their role as promoters is tightly linked to their hypomethylated state. If promoter function and transcription were responsible for keeping CGIs free of methylation, this would require all CGIs to be active during the wave of *de novo* methylation in the early blastocyst and during germ cell development (see 1.1.3) (reviewed in (Smith & Meissner, 2013)). Data from early studies show that this indeed might be the case (Daniels *et al*, 1997). For example the tissue-specific *α-globin* gene that possesses a CGI promoter is transcribed in the embryo, whereas the related *β-globin* gene that lacks a CGI is silent. In this and other reports highly tissue specific CGIs were demonstrated to be active in early development (Daniels *et al*, 1995; 1997). Recent large-scale findings strengthen this hypothesis by showing that 90% of genes with CGI promoters are expressed in the early embryo or testis (Sequeira-Mendes *et al*, 2009). Another report described that global transcription is a hallmark of embryonic stem cells and that the transcriptionally hyperactive ES cell genome undergoes large-scale silencing as the cells differentiate (Efroni *et al*, 2008).

Interestingly, CGI promoters seem to attract RNA pol II regardless of their transcriptional activity status. Also it has been shown that RNA pol II at CGI promoters engages in short bidirectional abortive transcripts, even at otherwise inactive genes (Core *et al*, 2008; Kanhere *et al*, 2010). These non-productive transcripts might play an important role in keeping CGIs free of methylation. One possible mechanism might be that RNA pol II is involved in the recruitment of H3K4 HMTs, which methylate lysine 4 on histone 3 (Guenther *et al*, 2007; Lee & Skalnik, 2008). H3K4me3 prevents chromatin binding of DNMT3L, a partner of DNMT3A and 3B and therefore interferes with DNA methylation. (Ooi *et al*, 2007). Additionally it has been shown that the ADD domain of the two *de novo* DNMTs fails to interact with H3K4me3 and that chromatin substrates with unmodified H3 tail are methylated more efficiently by DNTM3A and DNMT3A/3L complexes than chromatin trimethylated at H3K4 (Zhang *et al*, 2010b). Alternatively, H3K4me3 can be established at CGIs by Cfp1 in the absence of transcription (Thomson *et al*, 2010) (see 1.2.4 and 1.2.5). In addition, binding of the H3K4 HMTs of the MLL family may protect promoters of developmental genes from DNA methylation, and this is also likely to be instructed through their CxxC domains (Erfurth *et al*, 2008).

The histone variant H2A.Z, which is preferentially found at TSSs, is another potential candidate for preventing DNA methylation at CGIs. It has been shown that H2A.Z is strongly anti-correlated with DNA methylation in *Arabidopsis thaliana* (Zilberman *et al*,

2008), and that H2A.Z and DNA methylation are mutually exclusive at TSSs in mouse B-cells (Conerly *et al*, 2010). Although these observations are just correlative it seems plausible that the histone variant H2A.Z plays a role in inhibiting DNMTs.

The observation that many CGIs co-localize with origins of replication (ORIs) led to the speculation that replication might lead to the local exclusion of DNMTs (Delgado *et al*, 1998; Antequera & Bird, 1999). In ES cells more than 80% of CGIs co-localize with ORIs but a causal relationship between ORI function and CGIs has not been established (Mouse Genome Sequencing Consortium *et al*, 2002).

The majority of unmethylated CGI promoters in the human genome show significant strand asymmetry in the distribution of guanine and cytosines in the immediate vicinity of TSSs, whereas GC content and CpG frequency are evenly distributed. This asymmetric distribution of Gs and Cs, also known as "GC skew", allows the formation of R-loop structures, which occur when the newly transcribed G-rich RNA strand re-anneals back to its template C-rich DNA strand. The non-template G-rich DNA strand is then forced into an unpaired single-strand conformation. It has been proposed that the R-loop formation might inhibit *de novo* DNMT recruitment, thereby contributing to the hypomethylated state at CGIs (Ginno *et al*, 2012). How this inhibition of DNMTs is mechanistically happening is not fully understood, but it has been suggested that R-loops present inappropriate substrates for DNMT3B or that they may contribute to the recruitment of H3K4 HMTs by the single stranded DNA (Ginno *et al*, 2012). Interestingly, the results presented in this paper indicate that transcription might not be enough to protect a CGI from *de novo* methylation, since a construct orientated in a way that did not allow R-loop formation became methylated.

It has long been proposed that an alternative way of protecting CGIs from DNA methylation is the active removal of DNA methylation by demethylases. However, this hypothesis could not be proven due to the difficulties in deciphering DNA demethylation pathways. The recent discovery that the three members of the TET protein family, which also possess a CxxC domain, can convert 5mC into 5hmC has provided a potential mechanism leading to DNA demethylation (Tahiliani *et al*, 2009; Ito *et al*, 2010). TET1 preferentially locates to CGIs in mouse ES cells (Ficz *et al*, 2011; Wu *et al*, 2011a) and its depletion results in increased CpG methylation at CGIs (Wu *et al*, 2011b). This led to the hypothesis that one function of TET1 might be to erase methylation marks that are sporadically put at CGIs, thereby contributing to their hypomethylated state.

The above section discussed how CGIs might be protected from *de novo* methylation during early development. Since it has been shown that CGIs can become methylated at later stages, after early embryogenesis (Mohn *et al*, 2008; Meissner *et al*, 2008) the question remains how the majority of CGIs escape this methylation at stages when ubiquitous transcription is shut down and more restrictive expression patterns are established. As described in the section "targeting *de novo* DNA methylation" (see 1.1.4), DNMTs may be specifically recruited to their target genes after the initial wave of global *de novo* methylation. This would mean that CGIs are not targets of DNA methylation by default but can become methylated in certain situations that require concerted action of DNMTs and other histone-modifying enzymes (see 1.2.3).

1.2.3. CpG island methylation

Although the majority of CGIs are hypomethylated, a small subset (~ 2-5%) of promoter-associated CGIs shows a high level of DNA methylation (Eckhardt *et al*, 2006; Weber *et al*, 2007), which is associated with stable promoter silencing (Bird, 2002). Contrary to earlier assumptions that CGI methylation only occurs during specialized situations like X-inactivation and genomic imprinting, recent studies highlight the fact that CGIs can become methylated during normal development, often in a tissue-specific manner (Illingworth *et al*, 2008), and that aberrant DNA methylation is a hallmark of many cancers (Baylin & Herman, 2000).

In mammals, the expression level of genes from the X-chromosome in females, which possess two X-chromosomes, and males that have only one X-chromosome, needs to be equalised. This is accomplished by a process called X-chromosome inactivation whereby one female X-chromosome is randomly inactivated (reviewed in (Augui *et al*, 2011)). The non-coding RNA (ncRNA) *Xist* plays a pivotal role in this process by coating the X-chromosome that is to become silenced (Penny *et al*, 1996). Furthermore, *Xist* is regulated, at least in part, by differential DNA methylation of its promoter region, which is hypomethylated on the inactive X (Xi) and hypermethylated on the active X (Xa) respectively in somatic cells (Norris *et al*, 1994). *Xist* coating is followed by the establishment of a repressive chromatin state involving H3K9me3 (Heard *et al*, 2001), Polycomb recruitment and subsequent H3K27me3 of the X-chromosome (Silva *et al*, 2003). DNA methylation of CGIs on the Xi is essential for maintaining the silent stage of genes but is not the initiating event in gene silencing. This is supported by the observation that Xi

genes are already repressed prior to the acquisition of DNA methylation at these sites (Payer & Lee, 2008).

CGI methylation also has a well-characterized role in genomic imprinting where a subset of mammalian genes is expressed from only one of the two sister chromosomes depending on whether they are the maternally or paternally inherited copy. More than 80 genes are controlled by genomic imprinting, the majority of which are found in clusters. These clusters contain several imprinted genes and ncRNAs that are controlled by differential DNA methylation of imprinting control regions (ICR). For example, the ncRNA *Airn* is paternally expressed from the differentially methylated intragenic CGI of murine *Igf2r,* and represses transcription of several genes *in cis*, thus restricting their expression to the maternal allele (Wutz *et al*, 1997; Sleutels *et al*, 2002).

Also cases of tissue-specific differential methylation were described at a small but significant number of CGIs, which could play a role in controlling tissue-specific transcription (Schilling & Rehli, 2007; Illingworth *et al*, 2008; Eckhardt *et al*, 2006). Methylated CGIs have been characterized that associate with the promoter regions of germ line specific genes which are silenced in somatic tissues. Genes of the MAGE (Melanoma Antigen Encoding Genes) family are silenced during embryogenesis primarily through the methylation of CpG-rich promoter sequences (De Smet *et al*, 1999). During differentiation from ES cells into neurons a number of CGIs acquire methylation, the majority of which are already silenced in ES cells, providing further evidence that silencing precedes methylation (Mohn *et al*, 2008). Additionally, differential CGI methylation can occur within different somatic cell types, although these are relatively rare compared with differences between germline and somatic CGIs (Meissner *et al*, 2008). Genes involved in developmental processes such as members of the *Pax6* or *Hox* family have been found to exhibit cell type specific DNA methylation at CGIs (Illingworth *et al*, 2008). It has been shown that promoters with a low CpG content are more likely to become hypermethylated than CpG-rich promoters (Weber *et al*, 2007; Illingworth *et al*, 2010). Whereas promoter CGI methylation is a relatively rare event (~3%), orphan CGIs are methylated much more frequently (~15%). Especially intragenic CGIs are prone to gain methylation (20-30%) (Maunakea *et al*, 2010; Deaton *et al*, 2011; Illingworth *et al*, 2010). Alternative promoter usage or expression of ncRNAs might be regulated by intragenic DNA methylation (Dinger *et al*, 2008). It is also conceivable that regulation of splicing is influenced by intragenic CGI

methylation (Kornblihtt, 2006). More work will be needed to decipher the exact role of non-promoter CGI DNA methylation.

Many human cancers are connected with the aberrant hypermethylation of promoter CGIs. Epigenetic silencing of tumor suppressor genes is associated with improper promoter methylation (reviewed in (Baylin & Jones, 2012)). Initially it was thought that aberrant DNA methylation in cancers happens randomly, but by using genome-wide approaches it became apparent that certain CGIs become specifically targeted for *de novo* DNA methylation (Keshet *et al*, 2006). It is interesting to note that many of these *de novo* targets are CGIs marked by the repressive Polycomb complex (Ohm *et al*, 2007; Schlesinger *et al*, 2006). As with normal DNA methylation, it is possible that a HMT, in this case EZH2, is responsible for recruiting DNMTs to loci of erroneous DNA methylation (Viré *et al*, 2006; O'Hagan *et al*, 2011). Additionally, human cancers are often characterized by genome-wide DNA demethylation resulting in genome instability, aneuploidy or chromosomal rearrangements (Rodriguez *et al*, 2006).

1.2.4. CpG islands and chromatin

One function of CpG islands seems to be to allow the formation of a permissive chromatin environment that facilitates transcription factor binding and consequently transcription. Early studies have suggested that CGIs are made of non-nucleosomal DNA that is absent from bulk genomic fractions and that CGIs are depleted of the linker histone H1, which is considered repressive to transcription (Tazi & Bird, 1990). More recent work found that, unlike other promoters, CGI-containing promoters do not require ATP-dependent chromatin remodelling complexes to be activated but are intrinsically accessible (Ramirez-Carrozzi *et al*, 2009). Consistently, CGIs in macrophages showed a reduced density of histone 3 even in the un-induced state and *in vitro* chromatin assembly assays showed that CGIs are more reluctant to form nucleosomes than other genomic regions (Ramirez-Carrozzi *et al*, 2009). Another recent study analysed the influence of GC-content on nucleosome positioning and depletion and showed that CpG-content and CGI-width correlate with nucleosome depletion both *in vivo* and *in vitro* (Fenouil *et al*, 2012). As discussed in 1.2.2, a feature of CGI promoters is that they generally display paused RNA pol II (Core *et al*, 2008). It has been debated whether paused promoter transcription represents a cause or a consequence of open chromatin structures (Seila *et al*, 2009). Fenouil and colleagues argue that RNA pol II is not responsible for the formation of nucleosome depleted regions (NDR), but rather plays a role in the precise nucleosome positioning and in the enlargement of the NDR (Fenouil *et al*,

2012). The notion that unstable nucleosomes are an intrinsic feature of CGIs and not transcription dependent is strengthened by the discovery that CxxC domain containing proteins, many of which possess histone-modifying activities, specifically recognize unmethylated CpGs, thereby contributing to the creation of CGI chromatin architecture (see 1.2.5). These CxxC proteins are recruited by the underlying unmethylated CpG content and not by the transcriptional state of the associated gene (Blackledge *et al*, 2010; Thomson *et al*, 2010). The CxxC domain containing protein Cfp1 is thought to recruit the H3K4 HMT complex SET1A/1B to CGIs (Thomson *et al*, 2010). The resulting H3K4me3 mark is a general feature of CGIs and can act as a binding platform, via PHD or chromodomains, for proteins that support transcription initiation, such as TFIID, for the nucleosome remodeling factor NURF30 or histone acetyltransferase (HAT) complexes (Vermeulen *et al*, 2007; Wysocka *et al*, 2006; Saksouk *et al*, 2009). In contrast, the histone lysine demethylases KDM2A/2B that also possess a CxxC domain are demethylating H3K36me2 at CGIs (Blackledge *et al*, 2010). Experiments in yeast suggest that H3K36me2 is inhibitory to transcriptional initiation by acting as a binding site for a histone deacetylase (HDAC) complex via a chromodomain (Li *et al*, 2009). Together all these features seem to create a chromatin environment that is permissive to transcription initiation but that require additional transcription factors to engage in productive transcription. Apart from this permissive chromatin state, CGIs can also adopt a more repressive form. A subset of CGIs is marked by the repressive chromatin mark H3K27me3 that is established by the PRC2 complex (Bernstein *et al*, 2006). This class of promoters is referred to as "bivalent CGIs" as they retain the active chromatin mark H3K4me3 in addition to the repressive mark H3K27me3. Bivalency will be further discussed in section 1.3.4. In contrast, a small number of CGIs acquire DNA methylation during cell differentiation and this is associated with a constitutively repressed state (see 1.2.3).

1.2.5. CxxC containing proteins

The finding that MBD proteins specifically bind to methylated CpGs sparked the attempt to identify proteins that recognize unmethylated CpGs. Skalnik and colleagues identified a non-methylated CGBP (CpG binding protein) using a phage based ligand screen (Voo *et al*, 2000). CGBP was later re-named Cfp1 (CxxC finger protein 1 or CxxC1). Within this protein there is a cysteine-rich domain, the so-called Zinc finger CxxC domain (ZF-CxxC) that contains eight conserved cysteine residues and coordinates two Zn^{2+} ions. It was found that this ZF-CxxC domain binds specifically to a single nonmethylated CpG (Lee *et al*,

2001). This domain is present in a small group of proteins termed CxxC domain containing proteins that contain several histone-modifying enzymes (see Figure 1.2.5-1).

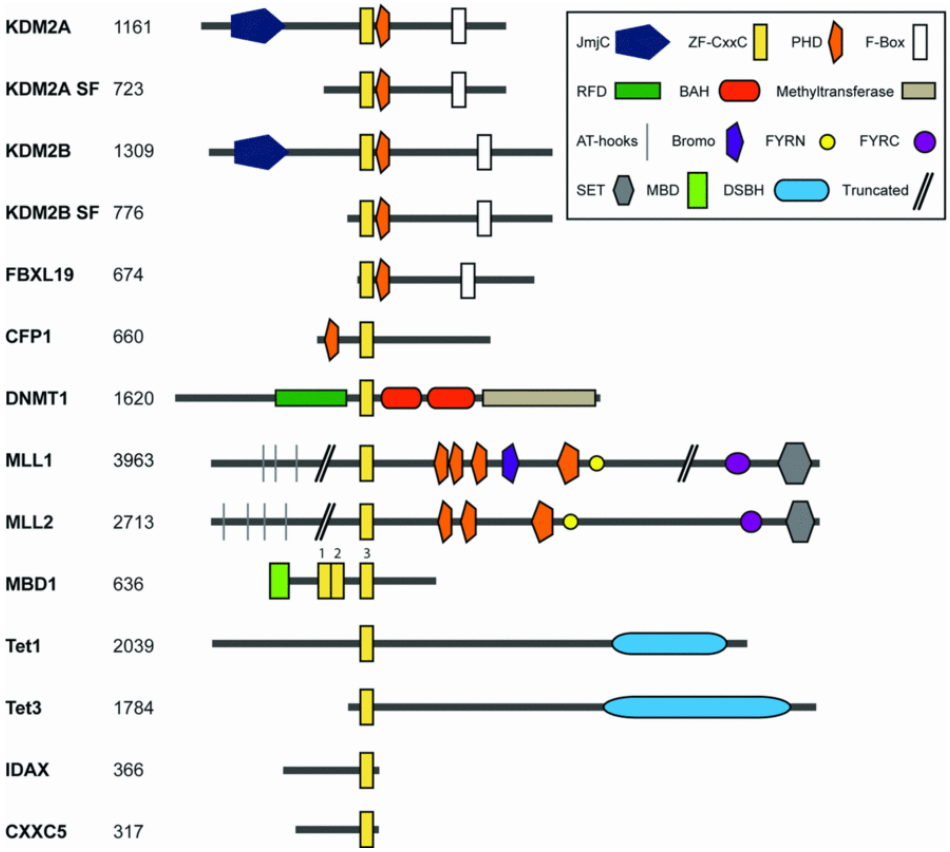


Figure 1.2.5-1 CxxC domain containing proteins.
An illustration of the domain architecture of mouse ZF-CxxC domain-containing proteins. The proteins are drawn to scale with the number of amino acids in the protein indicated on the left. The proteins are shown with the N-terminus on the left and all proteins are centered at the ZF-CxxC domain. In the case of KDM2A and KDM2B, alternative downstream promoters give rise to short forms of each protein (SF). Adapted from (Long *et al*, 2013).

ZF-CxxC domain containing proteins can be split into three subtypes depending on their sequence similarity. Whereas the eight cysteine residues are conserved in all three classes, only class 1 possesses an extended linker region that is located between the two cysteine-rich motifs and contains a highly conserved KFGG (Lysine-Phenylalaninie-Glycine-Glycine) motif. A KQ or RQ DNA-binding motif, which binds specifically to the CpG dinucleotide, is present in type 1 ZF-CxxC domains, is lost in class 2 proteins and contains an HQ motif in class 3 (see Figure 1.2.5-2 A). How this domain recognizes only unmethylated CpGs and not the methylated form, which only differs by the presence of a single methyl group, has only recently begun to emerge with the availability of exact structural models in both unbound

and DNA-associated states (Xu *et al*, 2012; 2011a; Allen *et al*, 2006; Song *et al*, 2010). When bound to DNA, the ZF-CxxC domain lies perpendicular to the DNA axis and interacts with the major groove via a DNA-binding loop. Regions flanking the ZF-CxxC domain interact with the minor groove. While the KFGG motif is not required for sequence specific DNA interaction, the KQ or RQ motifs form hydrogen bonds with the cytosines from both DNA strands and a guanine from one of the two strands, so that this DNA-binding loop penetrates the major groove (see Figure 1.2.5-2 B). This means that the ZF-CxxC domain clamps around the DNA, requiring access to both the major and the minor groove. When DNA is tightly wound around histone octamers access to both grooves is inhibited. Therefore the ZF-CxxC domain must bind to linker DNA between nucleosomes *in vivo* (Zhou *et al*, 2011). When a CpG dinucleotide becomes methylated, these hydrogen bonds cannot form any longer and DNA binding of the CxxC domain containing protein is inhibited (reviewed in (Long *et al*, 2013)). This indicates that the CxxC domain-mediated recognition of CGIs requires both the presence of non-methylated CpGs and accessible nucleosome depleted DNA. Two studies demonstrated recently on a genome-wide scale that CxxC domain containing proteins also function *in vivo* as CGI targeting molecules that contribute to the special chromatin structure found at CGIs (discussed in 1.2.4) (Blackledge *et al*, 2010; Thomson *et al*, 2010).

KDM2A/2B

KDM2A is a Jumonji C (JmjC) domain containing histone demethylase that was shown to preferentially catalyze the demethylation of H3K36me3 (Tsukada *et al*, 2006). It has been demonstrated that KDM2A recognizes unmethylated CpGs through its ZF-CxxC domain and that this binding is abrogated in a CxxC mutant protein or when methylated CpGs are present. Moreover, genome-wide ChIP-Seq data shows that KDM2A localizes to CGIs in mouse ES cells independently of the transcription status of the associated gene, indicating that KDM2A is recruited by unmethylated CpGs rather than by the transcriptional machinery (Blackledge *et al*, 2010). Additionally, KDM2A bound CGIs are depleted of H3K36me2, a histone mark that has been implicated in inhibiting transcription initiation in yeast, suggesting that KDM2A plays an active role in removing this mark from CGIs (Carrozza *et al*, 2005). The exact role of H3K36me2 in higher eukaryotes is still not fully understood, but it is possible that targeting of KDM2A to CGIs via its CxxC domain provides a mechanism of how a chromatin state at CGIs can be achieved that is permissive for transcription. KDM2B possesses a similar protein structure to KDM2A, displays H3K36me2 demethylase activity and localizes to CGIs (Farcas *et al*, 2012; Wu *et al*, 2013; He *et al*, 2013). Although

both KDM2A and KDM2B associate with CGIs, there is a specific subset of CGIs that is only bound by KDM2B. Upon further analysis of this subset it became apparent that many of those genes are Polycomb protein targets, suggesting that KDM2B might play a role in Polycomb-mediated transcriptional repression (Farcas *et al*, 2012). It was subsequently shown that KDM2B is a member of a variant PRC1 complex and hypothesized that KDM2B could recruit PRC1 to Polycomb targets in a PRC2-independent manner (discussed further in 1.3.3). The discrepancy that KDM2B is present at virtually all CGIs whereas PRC1 is only found at a small subset of CGIs was resolved by showing that the majority of CGIs is also occupied by PRC1, albeit to very low levels (Farcas *et al*, 2012).

FBXL19

F-box and leucine-rich repeat protein 19 (FBXL19) also possesses a ZF-CxxC domain but is devoid of the JmjC domain. Both, the KDM2A and KDM2B genes also have alternative transcription start sites giving rise to short forms of these proteins that are similar to FBXL19 (see Figure 1.2.5-1). The role of FBXL19 and the short forms of KDM2A and KDM2B remain poorly understood, but the presence of a presumably functional ZF-CxxC domain suggests that they might recognize CGIs.

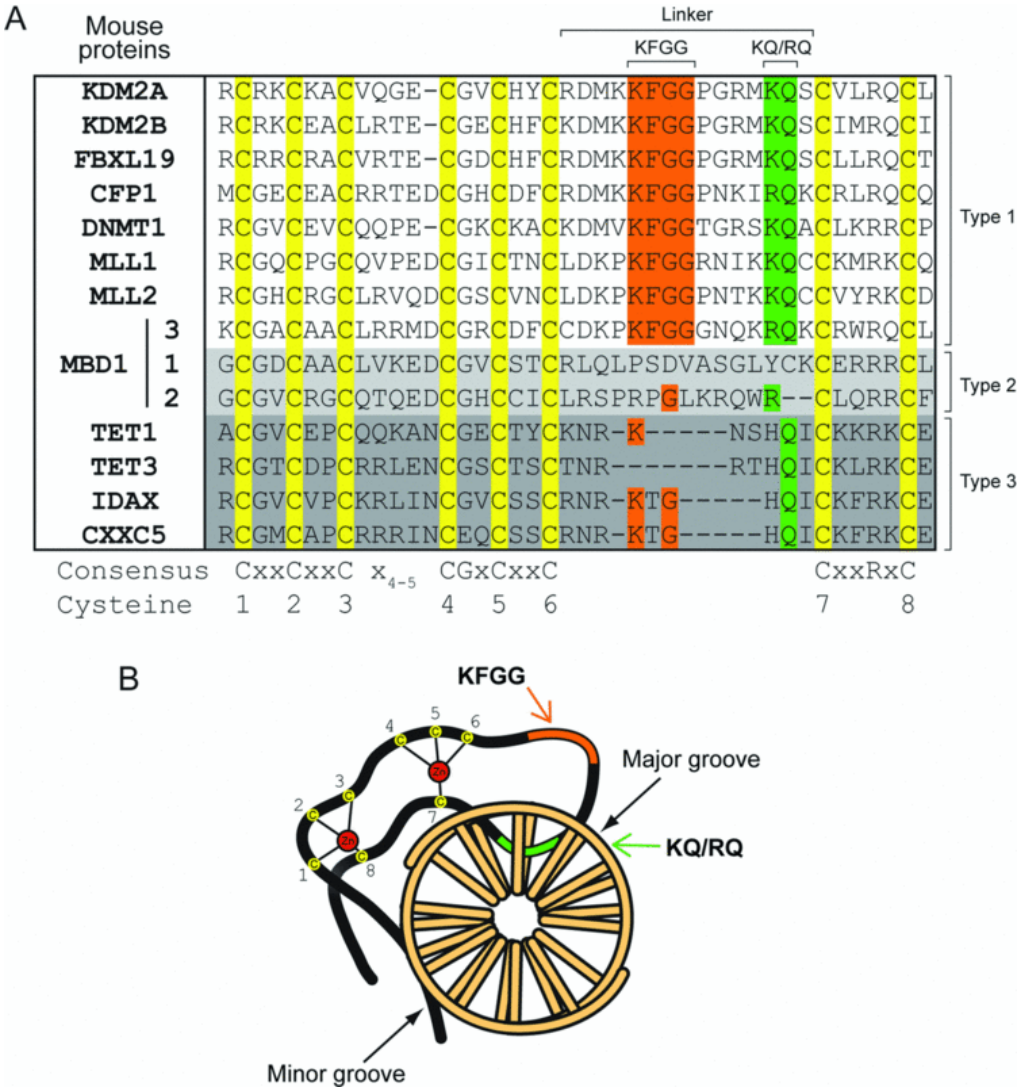


Figure 1.2.5-2 Sequence variation in ZF-CxxC domains & its interaction with DNA
(A) Sequence alignment of ZF-CxxC domains from mouse. ZF-CxxC domains can be split into three types depending on their sequence similarity, labeled as type-1, -2 and -3 on the right of the alignment. Conserved cysteins are highlighted in yellow. (B) A schematic of the ZF-CxxC domain highlighting the crescent structure and interaction with both the major and minor groove of DNA. Zn2+ ions (red) coordinate the eight cysteine residues (yellow). DNA is viewed down the double helix with bases shown as rods. Adapted from (Long *et al*, 2013).

Cfp1

Cfp1 (also CxxC1 or CGBP) is a ZF-CxxC domain containing protein that specifically binds to unmethylated CpGs. Mutation of either of two conserved cysteine residues in the CxxC domain to alanine results in ablation of DNA binding, suggesting that an intact CxxC domain is required for Cfp1 DNA binding activity (Voo *et al*, 2000; Lee *et al*, 2001). It also possesses a PHD domain and a SET1 interaction domain (SID), which is required for the interaction with the SET1A and SET1B H3K4 methyltransferase complexes (Lee & Skalnik,

2005; Lee *et al*, 2007a). Moreover, Cfp1 was found to interact with the maintenance DNA methyltransferase DNMT1 (Butler *et al*, 2008).

Cfp1 homologues have been identified in *Drosophila, C. elegans,* and both *S. cerevisiae* and *S. pombe* (Miller *et al*, 2001). SPP1 is the *Cfp1* homologue in yeast and is a component of the SET1/COMPASS histone H3K4 methyltransferase complex, the analogue of the mammalian H3K4 HMT complex. Interestingly, SPP1 lacks the CxxC domain in an organism that does not display CpG methylation. The SET1/COMPASS complex is the only H3K4 HMT in yeast (Briggs *et al*, 2001). In contrast, several mammalian H3K4 HMTs exist, each containing a component with a unique SET domain (see chapter 1.3.2).

Cfp1 null mice are not viable and die during early embryonic development, between day 3.5 and 6.5 dpc (days *postcoitum*). However, Cfp1 does not seem to be required for cell viability *per se* as *Cfp1* null blastocysts are viable and capable of inner cell mass and trophectoderm formation. Therefore, it is possible that Cfp1 is required for peri-implantation development, the developmental stage associated with global remodeling of chromatin structure and cytosine methylation patterns (Carlone & Skalnik, 2001). To further characterize Cfp1 function murine embryonic stem cells lacking Cfp1 were created. *Cfp1*$^{-/-}$ ES cells are viable but display an extended doubling time due to apoptosis and are not able to differentiate *in vitro*, consistent with the inability of *Cfp1* null embryos to gastrulate. They also display a 60-80% reduction in global cytosine methylation, including single copy genes, imprinted genes and repetitive elements, due to reduced activity of the maintenance DNA methyltransferase. In contrast, *de novo* methyltransferase activity remains normal in *Cfp1*$^{-/-}$ cells (Carlone *et al*, 2005). These deficiencies alone cannot account for the severe phenotype observed in *Cfp1* null embryos, as mouse embryos lacking DNMT1 die around 9 dpc whereas embryos lacking Cfp1 die much earlier. However, it is currently not clear how the lack of Cfp1 influences reduced DNMT1 activity.

Analysis of genome wide distribution of Cfp1 revealed that 80% of CGIs are associated with Cfp1 and this association does not depend on the activity state of the corresponding gene. Chromatin immunopreciptitation followed by genome-wide sequencing with antibodies against H3K4me3 showed that 90% of Cfp1-associated CGIs also possess H3K4me3 (Thomson *et al*, 2010). This result suggests that Cfp1 might be responsible for targeting the SET1 histone methyltransferase complex to CGIs via its CxxC domain. Consistently, knock down of *Cfp1* in NIH3T3 fibroblasts lead to a prominent reduction of H3K4me3 at CGIs

(Thomson *et al*, 2010). In support of this idea, insertion of promoter-less CpG-rich DNA can mediate *de novo* accumulation of H3K4me3 in ES cells (Thomson *et al*, 2010; Mendenhall *et al*, 2010). Genome-wide H3K4me3 ChIP-Seq data of *Cfp1*-/- mouse ES cells indicated that the around 50% of genes displayed a loss of H3K4me3; in line with the model that Cfp1 is important for targeting the SET1 complex to CGIs (Clouaire *et al*, 2012). The observed effect was greatest at promoters of active genes without an obvious effect on gene expression, implying that Cfp1 is not required in setting up basal H3K4me3 levels but is important for transcription-coupled deposition of this mark. Somewhat surprisingly, a substitution experiment with a mutated Cfp1 lacking a functional CxxC domain restored normal levels of H3K4me3 at affected genes. This result indicates that there must be an alternative mechanism that targets SET1 to basal H3K4me3 levels. Cfp1 contains redundant functional domains as introduction of either the amino-terminal half of Cfp1 that contains the CxxC domain or the carboxyl-terminal half that contains the SID domain is sufficient to correct the defects observed in the *Cfp1*-/- ES cells (Tate *et al*, 2009). However, no rescue was achieved with a full-length construct that contained point mutations in both the CxxC and SID domain, suggesting that either the DNA binding domain or the SET1 interaction domain is required for Cfp1 function (Tate *et al*, 2009). Cfp1 also possesses a PHD domain, which might be a possible candidate for targeting the SET1 complex to basal levels of H3K4me3, as it was recently shown that PHD domains are able to bind to H3K4me3 (Eberl *et al*, 2013). Different H3K4 methyltransferases such as MLL1 or MLL2, which also possess a ZF-CxxC domain, could be responsible for establishing these basal H3K4me3 levels.

In addition to reduced H3K4me3 levels at active genes, loss of Cfp1 results in the appearance of ectopic H3K4me3 peaks that localize to regulatory elements (Clouaire *et al*, 2012). Unlike H3K4me3 deficiency at promoters, ectopic peaks of H3K4me3 were not abolished by expression of the DNA-binding mutant of Cfp1, suggesting that the DNA binding activity conveyed by the CxxC domain is required to restrict the activity of the SET complex to bona fide targets and prevent mislocalization to other genomic regions. This is consistent with earlier findings that observed a global increase of H3K4me3 levels in *Cfp1*-/- ES cells, concomitant with reduced levels of heterochromatin, which suggested that Cfp1 restricts H3K4 methyltransferase activity to euchromatin (Tate *et al*, 2010).

MLL1/MLL2

The mixed lineage leukemia (MLL) H3K4 methyltransferase family comprises four large proteins, MLL1–MLL4, that form complexes that share a set of interaction partners with the

SETD1 complexes, including ASH2L, WDR5, RbBP5 and DPY-30 (see chapter 1.3.2.). MLL1 and MLL2 are closely related proteins that also contain a ZF-CxxC domain (see Figure 1.2.5-1), whereas MLL3 and MLL4 lack this domain. It has been shown that MLL1 and MLL2 bind to nonmethylated DNA *in vitro* via their CxxC domain (Birke *et al*, 2002; Bach *et al*, 2009), but it is not clear if that domain plays a similar role like Cfp1 and KDM2A/2B *in vivo*.

MLL1 is essential for normal development as it maintains the activity of developmentally important genes such as the *Hox* clusters. MLL1 knock out mice are embryonically lethal and show impaired segmental identity (Yu *et al*, 1995). MLL1 mostly localizes to gene promoters, which is reminiscent to the localization of Cfp1 and KDM2A/2B. However, unlike these proteins, which are present at CGIs regardless of the transcriptional state of the associated gene, MLL1 is restricted to a subset of CGIs that are found at active genes, suggesting that other mechanisms independent of ZF-CxxC-mediated targeting might play a role in MLL1 localization (Milne *et al*, 2005; Guenther *et al*, 2005). Chromosomal translocations that couple the N-terminus of MLL1, including the CxxC domain and an AT-hook, to the C-terminal domain of fusion partners have been implicated in driving aggressive adult and childhood leukemia (Meyer *et al*, 2009). Many of these fusion proteins are targeted aberrantly, leading to the erroneous expression of MLL1 target genes. MLL-fusion proteins with a mutant ZF-CxxC domain exhibit reduced transforming potential suggesting that the ZF-CxxC domain plays a crucial role in directing fusion proteins to genomic targets (Ayton *et al*, 2004; Milne *et al*, 2010).

Although MLL2 displays a very similar protein structure to MLL1 they have non-overlapping functions. Interestingly, MLL2 fusion proteins have not been implicated in promoting leukemia (Bach *et al*, 2009). It was shown that MLL2 is responsible for setting up H3K4me3 at bivalent gene promoters in mouse ES cells (Hu *et al*, 2013).

DNMT1

DNMT1 is the only maintenance DNA methyltransferase in mammals and it functions by associating with proliferating cell nuclear antigen (PCNA) at replication forks via its replication foci targeting sequence (RFTS). There, it copies the pre-existing DNA methylation pattern from the parental DNA strand to the newly replicated one. Consistent with its role to re-establish the symmetric methylation pattern after replication, it has been shown that its preferred substrate is hemimethylated DNA (see also chapter 1.1.2). Therefore

it might seem surprising that DNMT1 possesses a type 1 ZF-CxxC domain, which specifically recognizes unmethylated CpGs. A recent study shed light on this apparent discrepancy by demonstrating that this domain is required to restrict DNMT1 to its proper targets. This is achieved by an auto-inhibitory mechanism that prevents the catalytic domain from methylating DNA when the CxxC domain is bound to unmethylated CpGs (Song *et al*, 2010). When DNMT1 encounters a hemimethylated CpG, the CxxC domain cannot bind thus rendering the catalytic domain accessible. However, this structural study was based on a truncated version of DNMT1 that did not include the N-terminal RFTS. Full-length DNMT1 with mutations in the CxxC domain, which strongly impair DNA binding of the isolated CxxC domain, did not reduce the specificity of the enzyme, suggesting that the auto-inhibition model does not apply to full-length DNMT1 (Bashtrykov *et al*, 2012). It was also suggested that the RFTS domain can insert into the DNMT1 DNA-binding pocket and play an inhibitory role (Syeda *et al*, 2011) indicating that several mechanisms, dependent and independent of the CxxC domain, might play a role in restricting the catalytic activity of DNMT1 to hemimethylated CpGs.

MBD1

MBD1 is a transcriptional repressor protein that contains an MBD domain, with which it is able to specifically bind to methylated CpGs. Additionally, this protein possesses three CxxC domains, only one of which is functional and capable of binding to unmethylated CpGs *in vitro* (Clouaire *et al*, 2010). The presence of both, ZF-CxxC and MBDs, in MBD1 makes it possible that it could potentially read unmethylated and methylated CpG dinucleotides individually or in combination (Jørgensen *et al*, 2004). However, point mutations in the CxxC-3 domain which disrupt DNA binding *in vitro* did not affect the recruitment of MBD1, suggesting that functional MBD1 targeting can be achieved in the absence of the ZF-CxxC domain and that MBD1 is primarily a methylated CpG binding protein (Clouaire *et al*, 2010).

TET1 and TET3

TET1 and TET3 are members of the TET family of proteins that also comprises TET2. Apart from the catalytic double-stranded β-helix (DSBH) domain, they contain a type 3 ZF-CxxC domain. While TET2 does not possess a CxxC domain, its neighboring gene IDAX (CxxC4) has a ZF-CxxC domain that is very similar to TET1 and TET3. TET proteins have been implicated in the conversion of 5mC to 5hmC, and are thought to be part of the mammalian DNA demethylation system (see chapter 1.1.3) (Ito *et al*, 2010; Tahiliani *et al*, 2009). The

type 3 ZF-CxxC domain permits a more flexible way of DNA binding, allowing the binding of unmethylated Cs in any sequence context. This flexibility is achieved by a shortened linker region and a divergent DNA-binding loop, making the DNA-binding interface of the TET3 ZF-CxxC domain not as rigid as those found in type 1 CxxC domains (Xu *et al*, 2012). In contrast, it has been suggested that TET1 binds to methylated and unmethylated CpGs (Xu *et al*, 2011b; Zhang *et al*, 2010a) or that is does not show sequence specific DNA binding (Frauer *et al*, 2011). Despite these contradicting reports TET1 seems to localize to CGIs *in vivo* and shows a positive correlation with CpG density (Williams *et al*, 2011; Wu *et al*, 2011b). 5hmC seems to be least abundant at CpG-rich promoters with high levels of TET1. This initially surprising result might indicate that it is the function of TET1 to convert sporadic aberrant DNA methylation at CpG-rich promoters to 5hmC, which could be further processed to an unmethylated C. This hypothesis is supported by the fact that knock down of TET1 results in the accumulation of DNA methylation at specific CGIs (Xu *et al*, 2011b). Alternatively, it has been proposed that TET1 has a repressive role, additionally to its function of converting 5mC to 5hmC, that involves the direct recruitment of the SIN3A co-repressor complex (Williams *et al*, 2011).

1.3. Chromatin

In eukaryotic cells DNA is tightly packed into a nucleoprotein structure called chromatin. The basic unit of chromatin is the nucleosome that consists of 147bp of DNA wrapped around an octamer of 4 conserved histone proteins (2 copies of H3, H4, H2A and H2B) (Luger *et al*, 1997). The histone protein H1 interacts with the linker DNA separating two nucleosomes, thereby stabilizing the nucleosome structure and locking it in place (Happel & Doenecke, 2009). Strings of nucleosomes can be tightly packed and further condensed to form chromosomes. Chromatin can be modified in a variety of ways, enhancing or restricting accessibility of transcription factors and the transcription machinery. This can either be achieved by covalent modifications of histone tails, repositioning of nucleosomes by chromatin remodelling complexes or replacement of canonical histones with special histone variants. Chromatin modifications occur in a highly combinatorial and, sometimes, mutually exclusive fashion (Kouzarides, 2007; Ho & Crabtree, 2010; Mizuguchi *et al*, 2004).

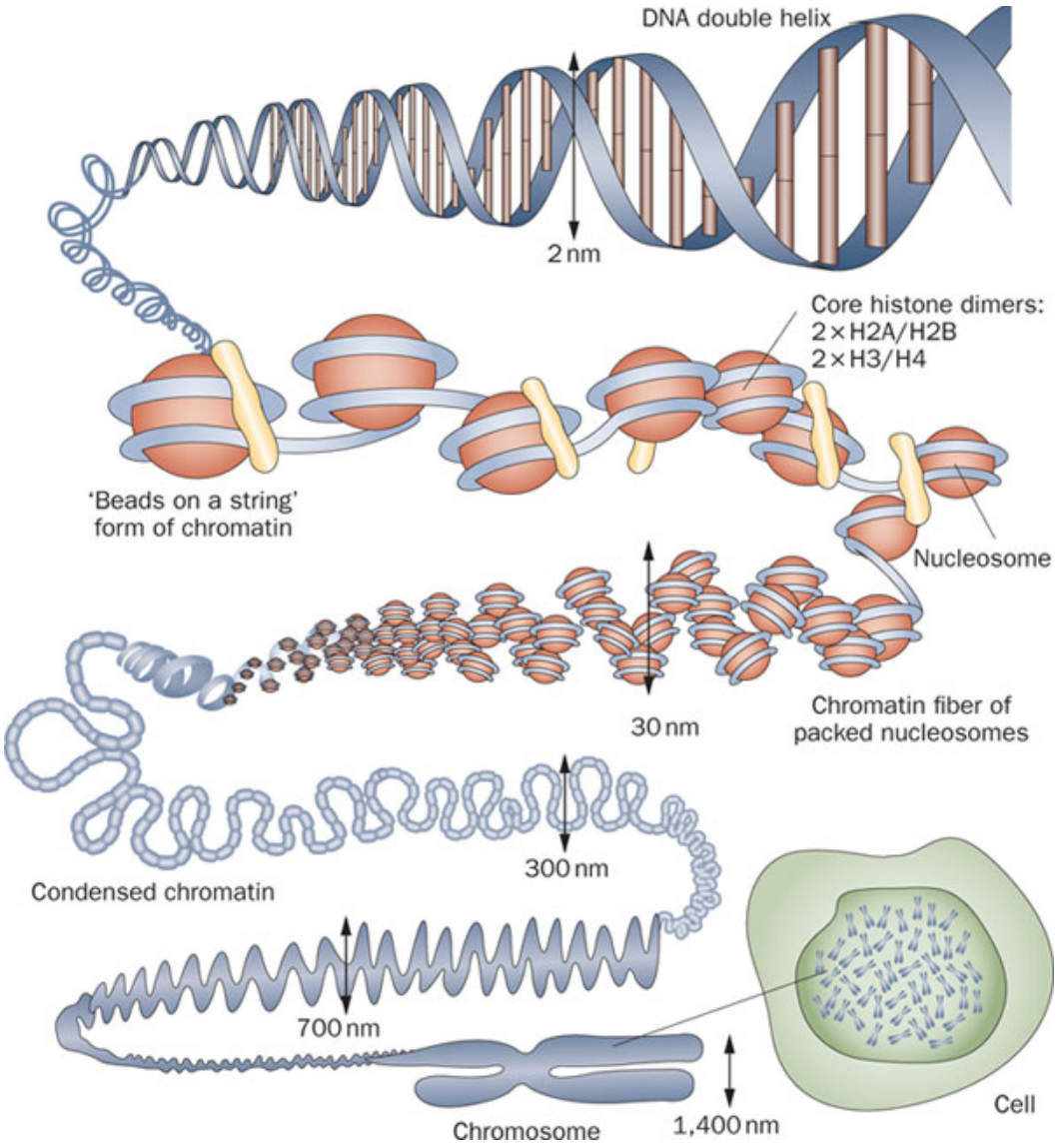


Figure 1.2.5-1 Chromatin organisation in the cell
The core particle of the nucleosomes is composed of 147bp of genomic DNA (2nm in diameter) wrapped around a histone octamer that consists of two copies of the major types of histones (H2A, H2B, H3 and H4). The yellow rod shape symbolizes the linker histone protein H1, which binds to DNA and histones, locking nucleosome in its place. Strings of nucleosomes can be tightly packed and further organized to form chromosomes. Adapted from (Tonna *et al*, 2010).

1.3.1. Histone modifications

Histones are globular proteins with a largely unstructured N-terminal tail that can be covalently modified in a number of ways. These modifications include methylation (me), phosphorylation (ph), acetylation (ac), ADP ribosylation (ar), ubiquitination (ub) and sumoylation (su), some of which correlate with active chromatin while others are associated with a repressed state (Kouzarides, 2007; Bernstein *et al*, 2007). Different chromatin states

are defined by combinatorial patterns of these histone modifications and each histone modification can induce or inhibit subsequent modifications. Histone modifications can have an effect on DNA-protein, protein-protein and nucleosome-nucleosome interactions by changing the charge of residues or providing binding platforms for a variety of proteins (Kouzarides, 2007). Initially the existence of a strict histone code was proposed that is recognized by transcription factors and dictates a transcriptional outcome (Strahl & Allis, 2000). However, an abundance of studies showed that the patterns of histone modifications are more complex than initially anticipated, revealing a multifaceted cross talk between histone modifications and context depending effects (Lee *et al*, 2010). To add to this complexity, the chromatin structure is also influenced by effector or "reader" proteins that recognize single or multiple histone modifications. The majority of histone modifications are reversible, the most prominent ones being histone deacetylases and histone demethylases, adding to the dynamics of chromatin.

Histone modifications such as, among many others, H3K4me2, H3K4me3 and H3/H4ac occur at gene promoters and are associated with transcriptional activity. H3K36me3 is found predominantly at gene bodies and correlates with transcriptional elongation. H3K9me2, H3K9me3 and H3K27me3, as well as hypoacetylation of H3 and H4, are implicated in transcriptional silencing. Acetylation has a positive influence on transcription by neutralising the positive charge of the histone tails, thereby destabilising DNA-histone interactions and facilitating chromatin decondensation (Luger *et al*, 1997). Methylation can have a positive or negative effect on transcription. Depending on the position of the methylation mark different proteins associate with histone lysine methylation. For example, H3K4me3 attracts proteins of the basal transcription machinery while inhibiting DNMTs (Vermeulen *et al*, 2007; Ooi *et al*, 2007). On the contrary, H3K9me3 is bound by HP1, promoting heterochromatin formation (Bannister *et al*, 2001), while H3K27me3 attracts Polycomb proteins (Schuettengruber *et al*, 2007). Genome-wide mapping of histone modifications led to the classification of distinct chromatin states depending on the combinatorial pattern of different histone modifications. In *Drosophila,* 5 or 9 different chromatin states have been predicted, respectively (Filion *et al*, 2010; Kharchenko *et al*, 2011). Another study suggested the existence of 15 different chromatin states in human cells, corresponding to repressed, poised and active promoters, strong and weak enhancers, putative insulators, transcribed regions, and large-scale repressed and inactive domains (Ernst *et al*, 2011).

Apart from histone modifications, nucleosome positioning can influence transcriptional outcome as DNA sequences necessary for transcriptional activation can be occluded by nucleosomes (Saha *et al*, 2006). In the 'ground state', nucleosome positioning is largely dictated by DNA sequence composition, which provides thermodynamically favorable localization sites (Ioshikhes *et al*, 2006). There are several families of chromatin remodeling complexes that physically disrupt nucleosome interactions through the use of ATP hydrolysis such as the SWI/SNF (Switch/Sucrose non-fermentable), CHD (chromodomain and helicase domain), ISWI (imitation SWI) and INO80 complexes (Ho & Crabtree, 2010). These complexes recognize specific histone modifications and are important for development and differentiation.

Additionally, the exchange of canonical histones with different variants is a way of modulating chromatin in response to cellular triggers such as DNA replication or DNA damage (Kamakaka & Biggins, 2005). The canonical histone H3.1 is replaced by the histone variant H3.3 in active chromatin in a replication-independent way and is thought to positively influence transcription through the expelling of H3K9me3 histones (Ahmad & Henikoff, 2002). Another H3 isoform, CENPA, is a component of centromere-specific nucleosomes and responsible for the formation of centromeric heterochromatin (Smith, 2002). The H2A variant H2A.Z is localized to the promoter proximal regions of active genes, and has been proposed to have a role in destabilizing chromatin to facilitate transcription (Schones *et al*, 2008). H2AX localizes throughout the genome and becomes rapidly phosphorylated in response to double strand breaks (Redon *et al*, 2002).

1.3.2. The Trithorax system

The activating Trithorax and the repressive Polycomb (see 1.3.3) system represent two classes of antagonistic factors that are highly conserved in metazoans. This system is essential for the maintenance of the expression pattern of developmentally important *Hox* genes and plays a role in stem cell identity, lineage specification, genomic imprinting and X-chromosome inactivation (Schuettengruber *et al*, 2007). It exerts its function by regulating and memorizing transcriptional activity and by maintaining a specific gene expression pattern, even when the initial cues are diminished, through histone modifications and remodelling. H3K4 methyltransferases together with SWI/SNF and NURF chromatin remodeling complexes are part of the activating Trithorax system.

The first H3K4 methyltransferase identified was Set1, from the yeast *Saccharomyces cerevisiae*, within a complex named COMPASS (Complex Proteins Associated with Set1) (Miller *et al*, 2001). Set1 is the only H3K4 methyltransferase in yeast and capable of mono-, di- and trimethylating H3K4 (Krogan *et al*, 2002). In mammals, on the contrary, at least 10 known or predicted H3K4 methyltransferases exist. H3K4 methyltransferases contain a Su(var)3-9/Enhancer of zeste/trithorax (SET) domain, which is responsible for catalyzing the addition of methyl groups to specific lysine residues. The SET domains in these proteins are either related to yeast Set1 as in the case of SET1A/B or to the *Drosophila* trithorax gene (Trx), as is in the case of the mixed lineage leukemia (MLL) family (MLL1-4). There are additional predicted H3K4 methyltransferases in mammals that are unrelated to yeast genes, such as ASH1, SMYD3, SET7/9 or MEISETZ (Ruthenburg *et al*, 2007). Why mammals have the need for several H3K4 methyltransferases when yeast only needs one is not entirely understood. However, evidence suggests that they are functionally non-redundant as they display a differential expression pattern and associate with different targets. Moreover, deletion of either H3K4 methyltransferase gives rise to distinct phenotypes in mice (Vastenhouw & Schier, 2012).

H3K4me3 is associated with active chromatin permissive to transcription. While H3K4me3 is solely enriched at active genes in yeast (Santos-Rosa *et al*, 2002; Ng *et al*, 2003), in mammals it is a feature of active and inactive genes at levels dependent on gene activity (Barski *et al*, 2007; Guenther *et al*, 2007). Several interactors of methylated H3K4, which bind to H3K4me3 via different domains such as WD40, Tudor or PHD, have been identified (Ruthenburg *et al*, 2007). For example it was shown that H3K4me3 recruits the basal transcription factor TFIID via the PHD finger of the TAF3 subunit (Vermeulen *et al*, 2007). Additionally, H3K4me3 could recruit nucleosome remodeling or histone acetyltransferase complexes that harbor subunits with PHD motifs (Wysocka *et al*, 2006). It has been suggested that H3K4me3 may modulate the kinetics of RNA pol II elongation to facilitate transcript processing (Terzi *et al*, 2011) or affect antisense transcription of regulatory RNAs (van Dijk *et al*, 2011). However, the exact relationship between H3K4me3 and gene transcription remains elusive. The MLL complexes can function as transcriptional activators, as loss of MLL leads to loss of H3K4 methylation and activation of transcription at *Hox* and other loci (Wang *et al*, 2009). Other studies have questioned initial views, that H3K4me3 is required for transcription. For example, depletion of DPY30, a component of mammalian H3K4 methyltransferases, leads to clear reduction of H3K4me3 in ES cells, yet had only minimal effects on transcription levels (Jiang *et al*, 2011). Another work examined the

effects of depletion of Cfp1, a unique component of the SET1A/B H3K4 methyltransferase complex, on H3K4me3 and gene expressions. Although a drastic loss of H3K4me3 at expressed CGI-associated genes was observed, this had only minimal consequences for transcription (Clouaire *et al*, 2012). Reduction of SET1A occupancy and histone H3K4me3 following WDR82 depletion, the other unique SET1A/B subunit, had no effect on RNA pol II occupancy or steady-state transcript levels for the examined housekeeping target genes (Lee & Skalnik, 2008). In yeast, deletion of the only H3K4 methyltransferase Set1 abolishes global H3K4 methylation levels but mutants are viable and do not display global transcriptome alterations (Briggs *et al*, 2001; Miller *et al*, 2001). The lack of effect on transcription might indicate that low levels of H3K4me3 might be enough to mediate its role in transcription, as a complete erasure of this mark is rarely reached.

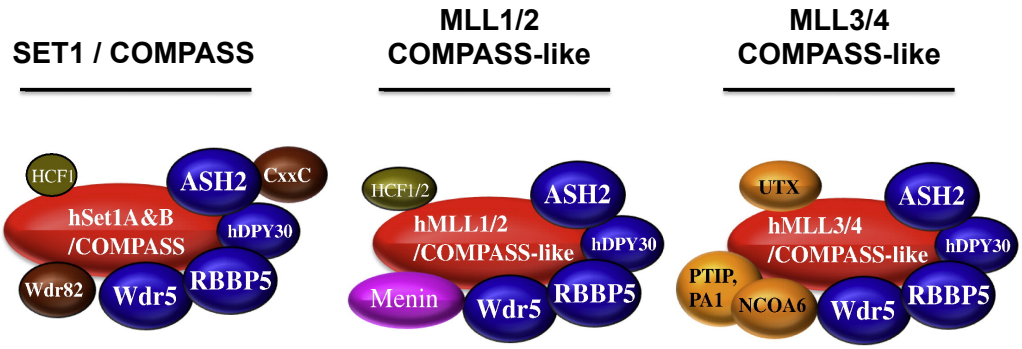


Figure 1.3.2-1 Mammalian H3K4 methyltransferases
Subunit composition of mammalian SET1A/B and MLL1-4 complexes. Each complex is capable of methylating histone H3 on its fourth lysine (H3K4). The known common subunits shared between yeast and mammalian complexes are shown in blue. WDR82 and CxxC, are found only in SET1A/B complexes. These subunits are shown in brown. MENIN, which is a subunit shared only in the MLL1/2 complexes is shown in lavender. MLL3/4 specific components are shown in gold. Adapted from (Eissenberg & Shilatifard, 2010).

Mammalian SET1A and SET1B, MLL1, MLL2, MLL3, and MLL4 form different complexes that contain unique subunits but also the common structural subunits WDR5, RBBP5 and ASH2 and DPY30 (see Figure 1.3.2-1) (Eissenberg & Shilatifard, 2010). The WDR5/RBBP5/ASH2 subcomplex associates with the MLL1 SET domain but can exist independently of the catalytic subunit, providing a structural platform that can associate with the SET domains of different MLL-family members (Dou *et al*, 2006). Structural studies of WDR5 reveal that it binds the H3 N-terminal tail in such a way that the K4 residue is fully exposed, a conformation that is ideal for further methylation (Ruthenburg *et al*, 2006).While ASH2 is not required for the stability of the MLL complex, it can regulate its catalytic

activity toward H3K4 trimethylation (Steward *et al*, 2006). Normal levels of DPY30 and RBBP5 are required for efficient ES cell differentiation (Jiang *et al*, 2011).

SET1A/1B

Human SET1A/1B functions most similarly to yeast COMPASS and mediates the bulk of H3K4me3 in mammalian cell extracts (Wu *et al*, 2008). Additionally to the core subunits ASH2, DPY30, WDR5 and RBBP5, SET1A and SET1B H3K4 methyltransferases contain two additional subunits, namely WDR82 and Cfp1 (CxxC1) (Wu *et al*, 2008; Lee & Skalnik, 2008; Lee *et al*, 2007a). WDR82 is required for the targeting of SET1A-mediated H3K4me3 near transcription start sites via tethering to RNA pol II (Lee & Skalnik, 2008). Additionally, WDR82 has a role in mediating the crosstalk between histone H2B ubiquitination and H3K4 methylation. In yeast, COMPASS can be recruited to actively transcribed genes through its interaction with the Paf1 complex and RNA pol II, which is sufficient for H3K4 monomethylation (Krogan *et al*, 2003). The Paf1 complex may also be involved in recruiting Rad6-Bre1 leading to ubiquitinated H2B (Wood *et al*, 2003). Cps35/WDR82 interacts with COMPASS in a histone H2B monoubiquitination-dependent manner, and this interaction on chromatin converts COMPASS to a trimethylation-competent complex (Lee *et al*, 2007b). Cfp1, the other unique subunit of SET1A/1B H3K4 methyltransferase complexes, possesses a CxxC domain that specifically recognizes unmethylated CpGs and might play a role in targeting the SET1 complex to CpG islands or restricting its activity to proper target genes (see 1.2.5 for more on Cfp1 and CxxC-domain containing proteins) (Tate *et al*, 2010; Thomson *et al*, 2010; Clouaire *et al*, 2012).

MLL1/2

MLL1 was originally discovered as the gene inducing human leukemia caused by chromosome band 11q23 translocations (Tkachuk *et al*, 1992). Both MLL1 and MLL2 include the unique subunit MENIN (Figure 1.3.2-1), which can act as a tumor suppressor and is required for localization of the complex to chromatin (Yang & Hua, 2007). MLL1 was found to be responsible for H3K4me3 of only a small subset of promoters in mouse embryonic fibroblasts (MEFs). Consistently, the loss of MLL1 does not alter bulk H3K4me3 but is required for the H3K4 methylation of developmental regulators such as *Hox* genes. In MLL1-/- MEFs these genes show decreased levels of RNA pol II and decreased gene expression (Wu *et al*, 2008; Wang *et al*, 2009). Recently, MLL2 was identified as the enzyme catalysing H3K4 trimethylation at bivalently marked promoters in embryonic stem cells (Hu *et al*, 2013).

MLL3/4

The MLL3 and MLL4 complexes are additionally associated with PTIP/PA1, NCOA6 and UTX, an H3K27 demethylase (Eissenberg & Shilatifard, 2010). They were recently identified as enhancer monomethylases (Herz *et al*, 2012). In this study the authors propose a model in which the transition from inactive/poised active enhancers is controlled by the combination of the H3K4 monomethylation activity of MLL3/MLL4 with the removal of H3K27me3 by UTX. As H3K27me3 is detrimental to H3K27ac, a hallmark of active enhancers, the removal of the H3K27me3 is a prerequisite for H3K27ac by CBP/p300 (Calo & Wysocka, 2013).

Recruitment of H3K4 methyltransferases

In flies, TrxG complexes bind to DNA elements called TrxG response elements (TREs), which often coincide with PcG response elements (PREs) (Schuettengruber *et al*, 2011). In mammals, no corresponding elements have been identified yet. It is likely that a combination of different recruiting mechanisms is responsible for the correct targeting of H3K4 methyltransferases. For example, MLL1 and MLL2 possess a CxxC domain via which they can be recruited to CGIs. CGIs also attract Cfp1, a subunit of the SET1A/1B complexes. Sequence-specific transcription factor binding plays an important role in H3K4me3 establishment. For example, the binding of transcription factors such as nuclear factor Y (NF-Y) and E2 to ASH2 induces its recruitment and mediates H3K4me3 at target promoters (Fossati *et al*, 2011). Therefore ASH2 has an important role in linking transcription factors to histone H3K4 methylation. Furthermore, ASH2 itself binds CG-rich DNA motifs, reinforcing the tethering of MLL complexes to their target chromatin (Sarvan *et al*, 2011). Additionally, H3K4 methyltransferases can be targeted by interaction with the polymerase associated factor 1 (PAF1) elongating complex, which directly interacts with sequences flanking the CxxC domain of MLL proteins (Muntean *et al*, 2010). A recent report provides evidence that long non-coding RNAs (lncRNAs) might function to activate transcription by recruiting H3K4 methyltransferases to chromatin. HOXA transcript at the distal tip (HOTTIP), a lncRNA from the 5′ end of the HOXA locus, was shown to interact with WDR5, targeting WDR5–MLL complexes across HOXA to induce H3K4me3 and gene activation (Wang *et al*, 2011). Alternatively, methyltransferases can be recruited by pre-existing chromatin marks. For example, MLL directly binds to di- and tri-methylated H3K4 via its PHD domain, which is necessary for its recruitment to target genes such as HOXA9 (Milne *et al*, 2010). The NURF complex can also read H3K4me3 via the PHD finger of one of its subunits, thereby linking H3K4me3-mediated gene activation with nucleosome

remodelling (Wysocka *et al*, 2006). Cross talk between different histone proteins also plays a role in chromatin establishment. For example, the human PAF complex recruits BRE1–RAD6, which ubiquitylates the H2B tail; this modification enhances H3K4 di- and tri-methylation by inducing the catalytic activity of SET1A/1B complexes (Kim & Buratowski, 2009).

In summary, different H3K4 methyltransferases show different functions and are required for H3K4me3. However, which role this histone mark plays with regard to transcription is less clear. H3K4me3 establishment could also provide a mechanism to control and balance H3K27me3, a mark written by Polycomb proteins (Smith & Shilatifard, 2010).

1.3.3. The Polycomb system

PcG proteins were originally identified in *Drosophila* as key mediators of heritable gene silencing, constituting a memory system for stable propagation of gene silencing through multiple cell generations (Simon & Kingston, 2009). In mammals, Polycomb repressive complexes (PRCs) represent one cellular mechanism to silence key developmental regulators. The Polycomb system consist of two main complexes, PRC2 that trimethylates H3K27 and PRC1-type complexes that ubiquitylate histone H2A and compact nucleosomes (see Figure 1.3.3-1) (Margueron & Reinberg, 2011; Simon & Kingston, 2009).

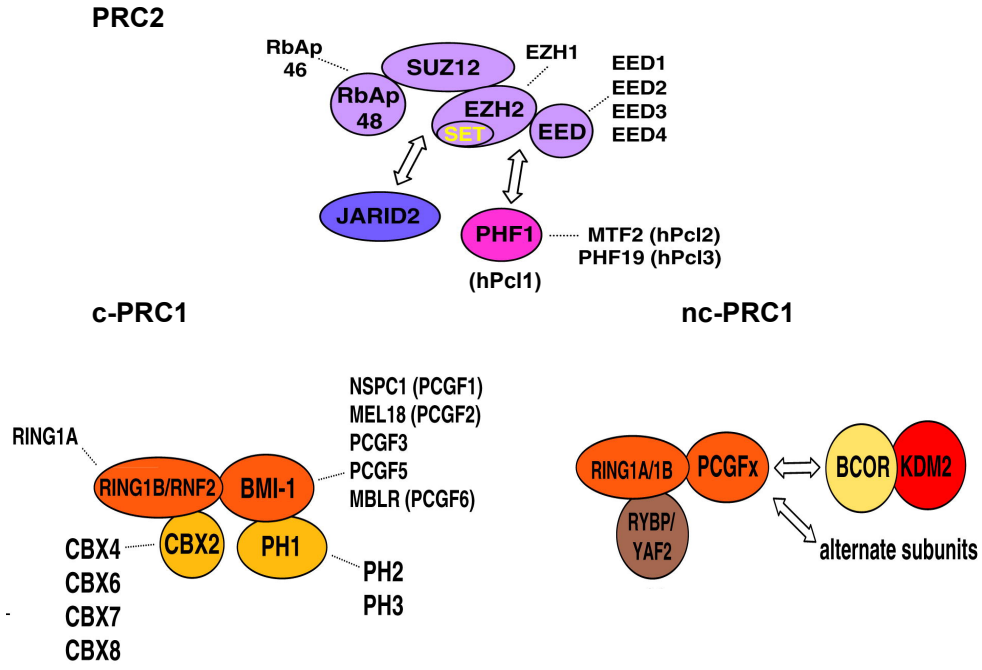


Figure 1.3.3-1 Composition of Polycomb complexes PRC2 and PRC1

Upper panel: Core subunits of PRC2 are in lavender, and arrows depict association of optional subunits. Dashed lines indicate alternative subunits derived from multiple gene copies or protein variants from a single gene. Lower Panel: Mammalian canonical PRC1 (c-PRC1) with four core subunits including CBX (a PC homolog). Non-canonical PRC1 (nc-PRC1) variants that contain KDM2 and/or RYBP subunits. In human PRC1 complexes, assembly of RYBP and CBX subunits are mutually exclusive. Adapted from (Simon & Kingston, 2013).

PRC2 complex

The core PRC2 complex, which is conserved from *Drosophila* to mammals, consist of 4 subunits: EZH1/2, EED, SUZ12 and RbAp46/48 (see Figure 1.3.3-1). EZH2 is the catalytic subunit that possesses a SET domain, which is responsible for di- and trimethylating H3K27 (Müller *et al*, 2002; Kuzmichev *et al*, 2002). However, EZH2 is catalytically inactive on its own and needs to be assembled with EED and SUZ12 to be active (Pasini *et al*, 2004; Cao & Zhang, 2004). Additionally to EZH2, which is the main H3K27 methyltransferase and only expressed in actively dividing cells, EZH1 is expressed in both dividing and differentiating cells but shows low methyltransferase activity. It is thought that while EZH2 is responsible for the majority of H3K27me2/3, EZH1 replaces H3K27me3 marks that were lost through histone exchange or demethylase action (Margueron *et al*, 2008). SUZ12 directly binds to EZH2, which promotes PRC2 assembly and plays a role in controlling PRC2 enzyme activity (Ketel *et al*, 2005). There is also evidence that SUZ12 mediates interactions with PRC2 cofactors such as JARID2 (Peng *et al*, 2009). Moreover, SUZ12 is the PRC2 subunit with the strongest affinity for a set of short ncRNAs emanating from the 5' ends of repressed target genes (Kanhere *et al*, 2010). EED adopts a donut-like β-propeller structure with a conserved aromatic cage in the donut hole that preferentially binds to H3K27me3 (Margueron *et al*, 2009). Thus, pre-existing H3K27me3 stimulates the activity of PRC2 and this positive feedback might maintain H3K27me3 levels in local chromatin during cell-cycle progression (Hansen *et al*, 2008). Polycomb-like (PCL) proteins comprise a major class of co-factors that influence PRC2 function. In mammals there are 3 types, PCL1-3, that function in skewing the catalytic activity of EZH2 towards the H3K27 trimethylation state (Sarma *et al*, 2008). PCLs also contribute to PRC2 recruitment in ES cells (Walker *et al*, 2010; Hunkapiller *et al*, 2012). JARID2, a histone demethylase of the Jumonji family, is another prominent PRC2-associated protein. (Pasini *et al*, 2010; Shen *et al*, 2009; Peng *et al*, 2009). However, its Jumonji domain lacks key catalytic residues, which seems to impede demethylase activity (Li *et al*, 2010). JARID2 impacts PRC2 recruitment, as knock out/down experiments show loss of PRC2 binding. In contrast, effects, stimulatory or inhibitory on H3K27me3 levels upon loss of JARID2, are controversial (Simon & Kingston, 2013).

PRC1 complexes

PRC1 complexes are more dynamic than PRC2, as multiple different complexes with a variety of subunits exist in mammals (see Figure 1.3.3-1). PRC1 complexes contain a central core consisting of RING1 and PCGF proteins that are responsible for catalysing the monoubiquitination of histone H2A at lysine 119 (H2AK119ub1) (Cao *et al*, 2005). Another non-enzymatic function of PRC1 is the compaction of chromatin (Francis *et al*, 2004). The canonical PRC1 complex additionally possesses CBX proteins, which recognize and bind to H3K27me3. This mechanism of hierarchical recruitment of PRC1 by H3K27me3 established by PRC2 was long thought to be the main recruiting mechanism of PRC1 (Simon & Kingston, 2009). However, recent studies have challenged this view by describing alternative PRC1 variants that lack CBX proteins and that do not require PRC2 activity to mediate H2AK119ub1 (Tavares *et al*, 2012; Morey *et al*, 2012b). These non-canonical PRC1 variants are characterised by the presences of RYBP or YAF2 instead of CBX and PHC proteins. Indeed, it has been shown that RYBP occupancy on chromatin is unchanged in EED null ES cells (Tavares *et al*, 2012). In contrast, CBX7 of the canonical PRC1 is completely absent from its target genes in those cells, consistent with the hierarchical model of recruitment (Morey *et al*, 2012b). Interestingly, RING1B, a component of both the canonical and non-canonical PRC1 complexes, is only partially lost from target genes in EED null ES cells (Tavares *et al*, 2012). This diversity among PRC1 complexes in ES cells poses the question whether they possess distinct or overlapping roles in gene repression. A recent report examined the genome-wide localization of the canonical and non-canonical PRC1 complexes in mouse ES cells and showed that both complexes co-occupy many of the same target genes, but also regulate genes independently of each other (Morey *et al*, 2012a).

Targeting Polycomb complexes

PcG proteins can occupy large genomic regions like the four *Hox* clusters, where Polycomb proteins are present for more than 100 kilo bases (kb). In contrast, other PcG target genes are isolated loci such as the *Ink4A/Arf* gene (Simon & Kingston, 2013). At least 10% of genes in mouse ES cells are targeted by Polycomb complexes (Mohn *et al*, 2008). How the PRC complexes are delivered to their genomic targets is well understood in *Drosophila*, where the recruitment of PcG complexes is mediated by specific DNA elements called Polycomb response elements (PREs). Sequence specific DNA binding proteins, such as PHO, recognize PREs and recruit Polycomb proteins (Ringrose & Paro, 2007). In mammals, however, the exact mechanism of recruitment remains elusive, partially due to heterogeneous complex composition, the fact that they are present at large genomic regions

and because many proteins, which mediate Polycomb recruitment in fly do not exist in mammals. Somme reports suggested the existence of a mammalian PRE and implied a role for YY1, the mammalian orthologue of the *Drosophila* protein PHO, in Polycomb recruitment (Lo *et al*, 2012; Sing *et al*, 2009). Yet genome-wide analysis in mammals did not show a clear overlap between YY1 and PcG target genes (Xi *et al*, 2007).

CGIs are thought to play a widespread role in PRC recruitment in mammals (see Figure 1.3.3-2). For example, genome-wide analysis correlated H3K27me3 presence with CpG-rich DNA and it was found that more than 97% of EZH2-bound sites in mouse ES cells correspond to CGIs (Ku *et al*, 2008). This finding was supported by showing that a bacterial artificial chromosome (BAC) construct containing a CpG island, or even bacterial DNA with similar characteristics, can recruit PRC2 components following their integration into mouse genomic sites (Mendenhall *et al*, 2010). Despite the striking overlap between CGIs and H3K27me3 presence, only a small subset of CGIs is marked by Polycomb proteins. It has been suggested that only those CGIs, which are devoid of any activating features recruit the PRC2 complex to establish the H3K27me3 mark. However, how exactly the PRC2 complex is recruited to CGIs remains unclear. No PRC2 associated CxxC containing proteins, which could mediate binding to unmethylated CpGs, have been identified. Other proteins, such as JARID2, that show binding preferences towards G+C rich DNA have been implicated in PRC2 recruitment (Li *et al*, 2010). Another study suggested that PCL3, a cofactor of PRC2, affects binding to CpG islands but no direct CpG binding has been shown (Hunkapiller *et al*, 2012). It is also possible that DNA methylation plays a role in regulating PRC2 as around 95% of PRC2 target genes are also bound by TET1, which is able to bind CGIs via its CxxC domain and is involved in the demethylation of methylated CpGs (see 1.1.3 and 1.2.5) (Wu *et al*, 2011b). TET1 activity is needed for full PRC2 recruitment in mouse ES cells, as knock down of TET1 leads to an impairment of EZH2 binding at a large fraction (around 70%) of PRC2 binding sites (Wu *et al*, 2011b). However, no direct interaction between TET1 and members of the PRC2 complex have been reported. As it has been shown that DNA methylation inhibits recruitment of Polycomb proteins (Wu *et al*, 2010a), TET1 could have a function in keeping CGIs methylation-free and therefore accessible to PRC2 proteins. CGIs are also involved in the recruitment of non-canonical PRC1 variants. The histone demethylase KDM2B/FBXL10 that is present in a complex called BCOR, is associated with non-canonical PRC1 complexes (Gao *et al*, 2012) and was found to bind to unmethylated CpGs via its CxxC domain (Farcas *et al*, 2012; Wu *et al*, 2013; He *et al*, 2013). Knockdown studies showed that KDM2B is required to target RING1B and H2AK119ub1 to specific loci

containing CGIs in mouse ES cells and that KDM2B is needed for full H2A ubiquitination activity in mammalian cells (Farcas *et al*, 2012; Wu *et al*, 2013; He *et al*, 2013).

Additional complexity to Polycomb targeting is added by the fact that several non-coding RNAs have been implicated in recruiting PRC complexes. As discussed in 1.2.3, the non-coding RNA (ncRNA) *Xist* plays a pivotal role in X-inactivation by coating the X-chromosome that is to become silenced (Penny *et al*, 1996). Coating with *Xist* RNA leads to heterochromatinization and the inactive X-chromosome becomes methylated at H3K27 in an *Xist*-dependent manner (Plath *et al*, 2003). The *Xist* RNA forms two long stem–loop structures, which interact with PRC2 *in vitro* (Zhao *et al*, 2008). A recent study that used a method to capture PRC2-bound transcripts in mouse ES cells found thousands of RNAs, among which was *Xist*, that associate specifically with PRC2, probably via EZH2 (Zhao *et al*, 2010). Besides *Xist,* other lncRNAs have been reported to bind to either PRC1 or PRC2. For example, the lncRNA HOTAIR, which is transcribed from the human HOXC gene cluster, was shown to regulate HOXD genes *in trans* by the recruitment of PRC2, followed by the H3K27me3 (Rinn *et al*, 2007; Tsai *et al*, 2010). The importance of HOTAIR in PcG function was called into question because deleting the *Hotair* gene in mice shows little effect on gene expression or H3K27me3 levels of *Hoxd* genes. Also, HOTAIR was found to show poor sequence-conservation between mouse and human (Schorderet & Duboule, 2011). Another study proposed that short ncRNAs play a role in Polycomb recruitment. The authors showed that short ncRNAs, which form a stem-loop structure and interact with PRC2 through SUZ12, are transcribed from Polycomb target gene in ES cells (Kanhere *et al*, 2010). Additionally, PRC1 was shown to interact with the locally encoded ANRIL lncRNA to regulate the INK4A/ARF locus (Yap *et al*, 2010). In summary, the sum of relatively weak interaction that are established by each subunit of the PRC2 complex could lead to recruitment of PRC2 as was suggested recently (Margueron & Reinberg, 2011).

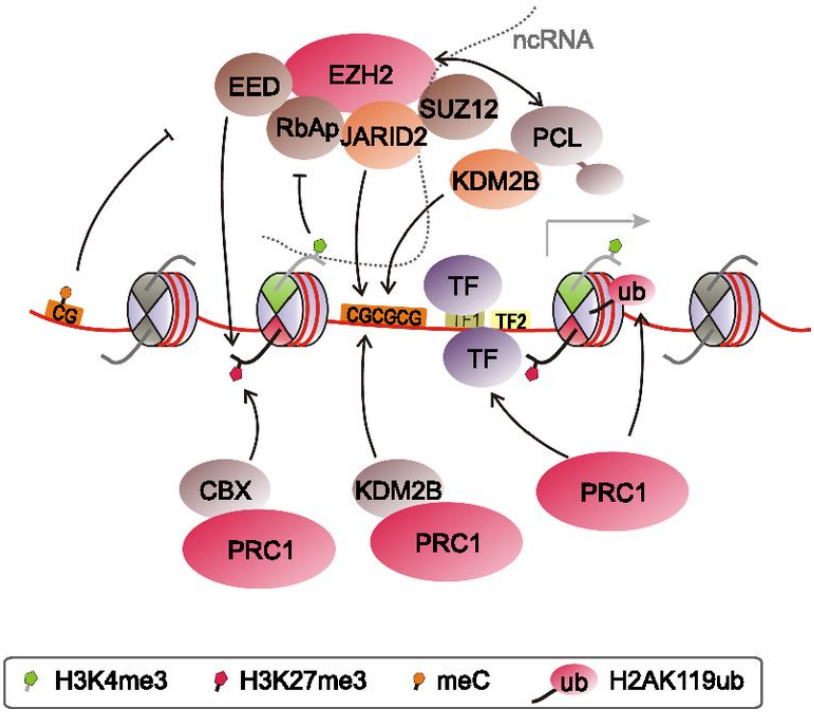


Figure 1.3.3-2 PRC2 and PRC1 recruitment to chromatin
Recruitment of PRC2 likely relies on interactions with DNA, histones, histone modifications, auxiliary proteins, and ncRNAs. It is proposed that it is the sum of weak interactions together with the absence of repelling factors that ultimately recruits PRC2. PRC1 can be targeted by the binding of subunit CBX to H3K27me3. PRC2-independent targeting of PRC1 is mediated by the recognition of unmethylated CGIs through the CxxC proteins KDM2B and through interaction with TFs. Adapted from (Voigt *et al*, 2013).

Effects of PRC2 and PRC1 on gene silencing

Despite many years of research, it is still not clear what the functional consequences of H3K27me3 are. It is still controversial if H3K27me3 plays an active role in gene silencing or if it is a secondary by-product of gene silencing rather than a cause (Henikoff & Shilatifard, 2011). Although it has been shown that inhibition of PRC2 does trigger gene re-activation (McCabe *et al*, 2012), a recent time course study implies that H3K27me3 accumulation occurs after the onset of silencing (Yuan *et al*, 2012). The initial view that H3K27me3 functions by targeting the PRC1 complex has been challenged as new finding reveal alternative recruitment mechanism for PRC1 (Tavares *et al*, 2012; Farcas *et al*, 2012; Wu *et al*, 2013). H3K27me3 could affect chromatin in many ways. For example, through the recruitment of other silencing factors, through antagonising activating H3K27Ac, impacts on nucleosome dynamics or by creating barriers for activating factors (Simon & Kingston, 2013). The effects of H2AK119ub1 on gene silencing have been investigated and it was found that two groups of genes exist, those that critically require H2AK119ub1 for silencing and those that display significant silencing without H2AK119ub1 (Eskeland *et al*, 2010;

Endoh *et al*, 2012). This indicates that, although H2AK119ub1 is important for gene silencing, alternative mechanism, likely chromatin compaction, play an additional role. It is not known how H2AK119ub1 influences gene silencing at a molecular level, but it might be that it effects chromatin structure or impedes transcription initiation or elongation (Simon & Kingston, 2013). Compaction of chromatin is the second well-studied function of the PRC1 complex, which works independently of H2AK119ub1, as it is unchanged when mouse RING1B is made catalytically inactive (Eskeland *et al*, 2010). The CBX subunit of PRC1 has been implicated in producing chromatin compaction (Grau *et al*, 2011). Interestingly, CBX proteins bind to H3K27me3 and promote PRC1 recruitment, while H3K27me3 is promoted by densely packed chromatin (Yuan *et al*, 2012). This provides a positive feed back loop between compacted chromatin and the chromatin mark that binds the PRC1 subunit responsible for compaction.

1.3.4. Bivalent genes

Genome-wide ChIP Seq studies of histone modifications in ES cells have generated location maps of H3K4me3 and H3K27me3. From these maps it became apparent that many gene promoters are associated with both H3K4me3 and H3K27me3 marks, which are referred to as "bivalent" genes (Mikkelsen *et al*, 2007; Bernstein *et al*, 2006; Ku *et al*, 2008; Pan *et al*, 2007). Interestingly, these bivalent chromatin domains are often found at developmentally important genes where they might contribute to the precise unfolding of gene expression programs during pluripotency and differentiation (Azuara *et al*, 2006; Mikkelsen *et al*, 2007). Despite the presence of the activating mark H3K4me3, bivalent genes are expressed only at very low levels (Bernstein *et al*, 2006). It has been suggested that the presence of H3K27me3 might contribute to the repression of lineage specific genes in ES cells, while the H3K4me3 mark serves to keep them poised for rapid activation upon differentiation (Bernstein *et al*, 2006; Ku *et al*, 2008). After differentiation into a specific cell type, continued association with H3K27me3 might maintain the repression of the majority of developmental control genes while only a specific subset of regulators is activated in a given lineage. Further studies confirmed the existence of bivalent genes in cell types other than mouse and human ES cells. Bivalent domains have been detected in haematopoietic stem cells, mouse embryonic fibroblasts (MEFs), neural progenitors and terminally differentiated neurons as well as in induced pluripotent stem cells (iPSCs) (Mohn *et al*, 2008; Mikkelsen *et al*, 2007; Cui *et al*, 2009; Guenther *et al*, 2010). These studies also revealed that, although the majority of bivalent domains are lost in differentiated cells, new bivalent domains are formed during differentiation (Mohn *et al*, 2008). H3K4me3 might also protect genes from

permanent silencing, for example by repelling transcriptional repressors or blocking DNA methylation (Fouse *et al*, 2008)(discussed in 1.3.2). The proposition that bivalent domains convey temporal and spatial control to the expression of lineage specific genes has proved to be popular and gained widespread attention.

Bivalent domains were initially predominantly observed in ES cells but it is less clear to which extent bivalent genes exist in the embryo and which role bivalency plays during development. Recent advances in ChIP technology that enables the analysis of chromatin states using low cell numbers led to the possibility of analyzing chromatin states in early embryos (Rugg-Gunn *et al*, 2010; Sachs *et al*, 2013). A study of mouse epiblast cells has also found putative bivalent domains but did not assess the simultaneous association of both chromatin marks in the same cell (Rugg-Gunn *et al*, 2010). Bivalent domains were further detected in the pluripotent inner cell mass of preimplantation murine embryos (Alder *et al*, 2010). Additional support for bivalent domains in developing embryos comes from studies in zebrafish, where bivalent domains were detectable by sequential ChIP at inactive genes, including the *Hox* clusters and other developmental transcription factors (Vastenhouw *et al*, 2010). Genome-wide analysis of mouse primordial germ cells (PGCs) revealed bivalent domains highly enriched at developmental regulatory genes (Sachs *et al*, 2013). While there is good evidence for the existence of bivalent domains in mammals and zebrafish, in other organisms the picture is less clear and the presence of bivalent gene promotes has been questioned. In one study very few bivalent domains were detected in gastrula stage *Xenopus* embryos (Akkers *et al*, 2009). Another study in *Drosophila* embryos has been unable to identify bivalent domains (Schuettengruber *et al*, 2009). It is conceivable that inherent differences in gene regulation between these species account for this apparent discrepancy. Alternatively, bivalent domains in *Xenopus* or *Drosophila* might arise in different developmental stages than the ones studied (Vastenhouw & Schier, 2012).

The above-described view of bivalent genes has been recently challenged by studies that suggest that bivalency is an artifact of heterogeneous cell populations and/or culturing conditions of ES cells (Hong *et al*, 2011; Marks *et al*, 2012). In order to establish bivalency at a given locus ChIP assays are often performed individually for each mark, which are unable to unequivocally establish the coexistence of both marks in the same cell. This led to the notion that bivalency could arise from the average of cells that carry either, but not both, marks at a given locus (Hong *et al*, 2011). Conventionally, ES cells are cultured in medium containing fetal calf serum. This requirement can be bypassed by using a specialized “2i

medium" that contains inhibitors of mitogen-activated protein kinase kinase (MEK) and glycogen synthase kinase-3 (GSK3), respectively. Cells grown in this medium are more homogeneous and exhibit tighter control of developmental gene expression. Notably, they display reduced levels of H3K27me3 at developmental promoters and fewer genes are classified as bivalent (Marks *et al*, 2012).

Principally, H3K4me3 and H3K27me3 marks at a bivalent domain can be present on adjacent nucleosomes, the same nucleosome, or even the same copy of H3 within a nucleosome. Sequential ChIP can be used to demonstrate that H3K4me3 and H3K27me3 are present on the same nucleosome or on neighbouring ones. Bivalency has been demonstrated on candidate genes using this method (Bernstein *et al*, 2006; Pan *et al*, 2007; Alder *et al*, 2010; Vastenhouw & Schier, 2012). However, to date no genome-wide sequential ChIP analysis has been published, which could interrogate the prevalence of true bivalent domains in a more general way. Using an approach combining mononucleosome ChIP with mass spectrometry-based quantitative profiling of histone modifications, a recent study showed that 15% of all H3 histones within H3K4me3-carrying mononucleosomes were marked with H3K27me3, arguing for a wide-spread presence of bivalently modified nucleosomes (Voigt *et al*, 2012). While it has been shown in these studies that bivalency within the same cell, at either the same or adjacent nucleosomes, does occur, questions remain regarding the exact position of the methylated histones within the nucleosome. It is unlikely that both H3K4me3 and H3K27me3 are found on the same H3 histone as it has been shown that PRC2 is inhibited by the active H3K4me3 mark (Schmitges *et al*, 2011). Moreover, mass spectrometry-based studies found that H3K4me3 and H3K27me3 do not coexist on individual histones in HeLa cells (Young *et al*, 2009). Accordingly, Voigt and colleagues found that many nucleosomes are asymmetrically modified, with H3K4me3 and H3K27me3 marks being present at opposite H3 tails (Voigt *et al*, 2012).

Function of bivalent domains

As mentioned above, the initial proposition that bivalent genes are kept in a primed state, which allows for either rapid activation or stable gene silencing upon differentiation while maintaining a reversible, silenced state in ES cells has received a lot of attention (Azuara *et al*, 2006; Bernstein *et al*, 2006). Genome wide studies later-on seemed to confirm this concept as resolution of bivalent genes and associated gene expression changes were observed upon differentiation (Mikkelsen *et al*, 2007; Pan *et al*, 2007; Zhao *et al*, 2007). However, from these studies it was not clear whether the observed effects are causative or

whether bivalency and its associated histone marks are a consequence of other factors such as the transcriptional status.

It was initially thought that H3K27me3-mediated repression of lineage regulators was essential for maintenance of ES cell pluripotency (Boyer *et al*, 2006; Lee *et al*, 2006). Indeed, loss of components of the PRC2 complex led to a decrease of H3K27me3 levels concomitant with partial de-repression of genes normally targeted by Polycomb (Shen *et al*, 2008; Pasini *et al*, 2007). However, despite the ectopic expression of transcription factors involved in lineage specification and a higher propensity to differentiate, cell viability and self renewal were not compromised in PRC2-deficient ES cells (Chamberlain *et al*, 2008; Leeb *et al*, 2010; Pasini *et al*, 2007). This could be explained by the fact that PRC1, which can be recruited to target genes independently of PRC2, is able to compensate for PRC2 loss. Indeed, a double knock out of RING1B and EED in ES cells resulted in more severe defects (Leeb *et al*, 2010). Many transcription factors required for differentiation are not expressed in ES cells. Therefore the effects observed from insufficient repression of lineage specific genes in Polycomb mutants might be underestimated because appropriate transcription factors are not present to increase aberrant expression. Additionally, one needs to bear in mind that disrupting Polycomb proteins does not only affect bivalent genes but also repetitive elements and telomeres, which are normal targets of Polycomb. While these studies demonstrated that ES cells are viable with strongly reduced H3K27me3 levels, it became apparent that this mark is required for proper differentiation. Although mutant ES cells can differentiate into ectoderm, mesoderm and endoderm, lineage regulators are not properly activated and ectopic activation of genes from alternative lineages might interfere with the execution of the proper developmental program (Pasini *et al*, 2007; Leeb *et al*, 2010; Chamberlain *et al*, 2008; Shen *et al*, 2008). This is in accordance with severe phenotypes, observed in mice lacking components of the Polycomb system, most of which result in post-implantation embryonic lethality (Vastenhouw & Schier, 2012).

With respect to H3K4me3, a recent study found that depletion of DPY-30, a core subunit of MLL histone methyltransferase complexes, results in a partial reduction of this activating mark, which is associated with improper activation of some lineage specific genes upon differentiation (Jiang *et al*, 2011). As observed with PcG depleted cells, depletion of DPY-30 did not affect ES cell-specific gene expression (Jiang *et al*, 2011). In contrast, reduction of H3K4me3 levels by depletion of WDR5, another subunit of MLL complexes, led to severe defects in ES cell maintenance and reduction in the expression of pluripotency genes (Ang *et*

al, 2011). Knock-out studies of members of MLL complexes support an *in vivo* role for H3K4me3 in transcription regulation and lineage specification and many mutants show early embryonic lethality (Vastenhouw & Schier, 2012). All these studies however did not test the proposed function of H3K4me3 in poising for the expression of embryonic genes and facilitating expression upon differentiation. One very recent study, which implicated MLL2 in the establishment of H3K4me3 at bivalent genes, did not detect substantial defects in rapid transcriptional induction after retinoic acid treatment of MLL2-depleted cells, questioning the poising model (Hu *et al*, 2013). As developmental promoters need to be protected from permanent silencing, through for example DNA methylation, mechanisms to keep them active need to be employed. H3K4me3 could play that role by preventing *de novo* DNA methylation (Ooi *et al*, 2007). Moreover, H3K4me3 interacts with factors of the transcription machinery such as TAF3 (Vermeulen *et al*, 2007). Such a permissive chromatin however can lead to spurious transcription that needs to be counter-balanced, presumably through the action of the Polycomb system. H3K27me3 might play a role in impeding deposition of H3K36me3 (Schmitges *et al*, 2011). Additionally, the ability of PCL3 to recognize H3K36me3 may be a trigger for increased PRC2 recruitment to dampen transcription (Cai *et al*, 2013). It is equally important to prevent PRC2 from excessive spreading. This can be achieved by a symmetrically H3K4 trimethylated nucleosome that inhibits PRC2 (Voigt *et al*, 2012). A recent study implicated a function for the transcriptional regulator UTF1 in balancing this system by limiting PRC2 recruitment and promoting mRNA degradation of aberrantly transcribed genes (Jia *et al*, 2012).

Upon differentiation, bivalent domains need to be resolved. Activating transcription factors together with H3K27 demethylases might shift the balance towards activation. In contrast, removal of the activating stimuli in concert with H3K4 demethylase activity might shift genes to a repressed state (Voigt *et al*, 2013). All this evidence suggests that bivalent domains function in the fine-tuning of gene expression during development.

1.4. PhD Objectives

The aim of this thesis is to explore the constraints on DNA sequence and genomic location that are required to impose both H3K4me3 and H3K27me3 at CGI sequences. It will be analysed whether DNA sequences with a high overall G+C content and a high CpG frequency are sufficient to establish novel marks of H3K4me3 and/or H3K27me3 independent of their location. This will be achieved by introducing promoter-less CGI-like sequences into so called “gene deserts”, which are sequences devoid of transcriptional activity. After analysing the chromatin state of these sequences it should become clear whether primary DNA sequences at CGIs directly influence chromatin structure. Additionally, the relative contribution of CpG frequency versus G+C content will be investigated. To this end sequences will be generated that either contain few CpGs but have a high G+C content or many CpGs with a low G+C content. This study will provide insights into the influence of CGIs on chromatin structure.

2. Material and Methods

2.1. Material and Reagents

2.1.1. Bacterial reagents for cloning and recombineering

LB medium:

Bacto tryptone (10g/l), Bacto yeast extract (5g/l), NaCl (10g/l), pH adjusted to 7.0 with NaOH. Bacto agar (20g/l) was added if making LB agar. Solutions were autoclaved prior to use. LB agar plates were stored inverted at 4°C and LB broth was stored at R/T.

SOC medium:

Bacto-tryptone (20g/l), Bacto-yeast extract (5g/l), NaCl (0.5g/l), 1M KCl (2.5 ml/l) and, pH adjusted to 7.0 with NaOH, autoclaved and 20 ml of sterile 1 M glucose were added before use.

Ampicillin stock solution:

50 mg/ml ampicillin in dH_2O. Filter sterilised (0.2μm filter) and stored at −20°C. Added to LB medium at a final concentration of 50μg/ml.

Kanamycin stock solution:

50 mg/ml kanamycin in dH_2O. Filter sterilised (0.2 μm filter) and stored at −20°C. Added to LB medium at a final concentration of 50 μg/ml for high copy plasmids and 15 μg/ml for low copy plasmids.

Chloramphenicol stock solution:

30 mg/ml chloramphenicol in 96% ethanol. Stored at −20°C protected from light..

Tetracycline stock solution

c = 10 mg/ml dissolved in 75% ethanol. Stored at −20°C protected from light. Added to LB medium at a final concentration of 3 μg/ml for low copy number plasmids.

Blasticidin stock solution

c = 10 mg/ml dissolved in H_2O. Stored at −20°C protected from light. Added to LB medium at a final concentration of 30 μg/ml for low copy number plasmids.

L-arabinose stock solution:

10% L-arabinose (Sigma A-3256) was prepared in ddH_2O, fresh or frozen in small aliquots at -20°C. 50 μl stock solution per 1.4 ml LB or SOC was used for induction of recombination protein expression from pRedET.

Blue/white selection plates:

LB plates containing the appropriate antibiotic were spread with 40µl X-gal (5-bromo-4-chloro-3-indolyl-β-D- galactopyranoside; 40mg/ml) and dried at 37°C. Prepared on day of use.

Orange G loading buffer (6×):

0.198% (w/v) orange G, 12% (w/v) Ficoll, 120 mM EDTA pH 8.0, 4.2% (w/v) SDS. Stored at −20 °C (long-term) or RT (short-term).

Proteinase K stock solution:

20 mg/ml proteinase K, 100 mM EDTA pH 7.5, 2% (w/v) SDS. Stored at −20 °C.

RNase A stock solution:

10 mg/ml RNase A in dH2O. Stored at −20°C.

DNA Sequencing Buffer (2.5x):

20 mM Tris HCl (pH8) and 5 mM MgCl2.

TAE electrophoresis buffer (1x):

40 mM Tris-acetate, 1 mM EDTA

TE buffer pH7.5:

10 mM Tris HCl pH 7.5, 1 mM EDTA

Hyb+ Buffer for Southern blot:

20xSSC, 50x Denhardts, 20% SDS, Dextran sulphate, ssDNA

Wash buffers for Southern blot:

2xSSC: 100ml SSC + 890ml H2O+ 10ml 20% SDS

1xSSC: 50ml SSC + 950 ml H2O + 10ml of 20% SDS

2.1.2. Cell culture reagents

Cell culture reagents were obtained from Gibco unless otherwise stated

2.1.3. ChIP reagents

ChIP buffers

Wash Buffer A

0.25% Triton X-100

10mM EDTA pH 8.0

0.5mM EGTA pH 7.5

10mM Hepes pH 7.5

Wash Buffer B

0.2M NaCl 1mM

EDTA pH 8.0

0.5mM EGTA pH 7.5

10mM Hepes pH 7.5

Lysis buffer

20% SDS
0.5M EDTA
1M Tris pH 8.1
50X PI tablet
1000x PMSF in isopropanol
100x Na Butyrate

Dilution buffer

5M NaCl
1M Tris pH 8.1
0.5M EDTA
10% Triton X-100
20% SDS

Wash Buffer 1

20mM Tris pH8

150mM NaCl

2mM EDTA

1% Triton X-100

0.1% SDS

Wash Buffer 2

20mM Tris pH8

500mM NaCl

2mM EDTA

1% Triton X-100

0.1% SDS

Antibody	Source & Catalogue number	Concentration
α-H3K4me3	Abcam-8580	3 µl
α-H3K4me1	Abcam-8895	5 µl
α-H3K27me3	Millipore-07-449	5 µl

α-H3K9/K14ac	Abcam 12179	5 µl
α-H3	Abcam 1791	3 µl
α-SUZ12	Abcam 12073-100	5 µl
α-RNA Pol II N20	Santa-Cruz -899	20 µl
α-RNA Pol II S5P	Abcam 5131	5 µl
α-RNA Pol II unphosphor.	Abcam 817	5 µl
α-GFP	Chromotec GFP-TRAP_A gta-20	5 µl
α-IgG	Invitrogen 10500C	2 µl

Table 1 List of antibodies used for ChIP

2.1.4. Primers used in this study

Mecp2-eGFP	**Forward Primer**	**Reverse Primer**
Irak TSS	AATGGGAGAACCGAGGTCTG	TTTGGCTGGCTCTCAAACTG
Mecp2 5	GATGATCCACAGGCAGCAAC	TCAAAGAAGAGGCCCCAGTG
Mecp2 7	CTCAAAAGAGCCCAGCTCTT	GGCAGCTGCAGTGCTGAACC
Mecp2 9	GACAGGGCATCAATGGCACC	CCAGACAAGCTGTTGACCAG
GFP A	AAGGGCGAGGAGCTGTTCA	CCGGTGGTGCAGATGAACTT
GFP B	GCGCACCATCTTCTTCAAGG	TCTGCTTGTCGGCCATGATA
GFP C	TCAAGATCCGCCACAACATC	ATGTGATCGCGCTTCTCGTT
Mecp2 14	GATGATGGTGCTCCTTCTTA	TCCACCCTTGGTGAGAAAAG
Mecp2 15	CAGGGCTCTTCTCCAGGACC	GAGCAGAAACCACCTAAGAA
Mecp2 16	GATCTGAAAACAGAGGACCT	GATTAGGTCTGGGTACCAAC
Nanog PuroGFP	**Forward Primer**	**Reverse Primer**
Nanog 11	CAGTGATTTGGAGGTGAATT	CCCAGATGTTGCGTAAGTC
Nanog 12	CAAACCTAGGACTTAGAACA	GCACCTCACTGTCTCCAAA
Nanog 13	GGGTCTTGGTATACACTGT	AGGTTGGCCTTGAACTTATT
Puro A	TCTGGACCACGCCGGAGA	CCAGGAGGCCTTCCATCTGT
GFP B	AAGGGCGAGGAGCTGTTCA	CCGGTGGTGCAGATGAACTT
GFP C	TCAAGATCCGCCACAACATC	ATGTGATCGCGCTTCTCGTT
22	GCCTTAGTCAACTGACATCT	CTGATGCCAAAGGACAGAAA
23	GCCTTAGTCAACTGACATCT	GGACAGTTGCCAGACAGAGG
Gene desert 1	**Forward Primer**	**Reverse Primer**
Gd1_1=ChIP E19_1	AGCAATGGCAGAATGAGAGG	TCAAACTGGACACAGGGACA
Gd1_2=ChIP E19_2	GCACTGGGTAGCCATTTCAG	TTCTTCGTTCTCCACGATCC
Gd1_3=ChIP E19_3	GGATCGTGGAGAACGAAGAA	TGTGTGCGTGTTTGTCTCTGT
Gd1_7=ChIP E19_5	TTATTACGCCATGGCTCCAG	CACAATTTGGGGCATGACAG
Gd1_8=ChIP E19_7	GCAGAGCATGCTGTTGTTGG	CCAGGTAGGTCTTTGGTGCTG
Gd1_9=ChIP E19_8	CCAGCTGCTGACAGGGAATC	TTGAATATGGCTTCTCTCTTTCCAC
Gene desert 2	**Forward Primer**	**Reverse Primer**

Gd2_1=ChIP L19_1	TGATGGTGGCATATTGCTTTC	CAATGCTCCATTTCCCTCTCA
Gd2_2=ChIP L19_3	GGAAGGGTCACTGTGTCAAGC	GAGGCTGACCTTTCTCCTCCA
Gd2_3=ChIP L19_6	TGACTCTCTGCCATGATGTGG	CAGGAAAGCTGTTCCCACGA
Gd2_4=ChIP L19_7	GCATTTCAGGCTGAGGTATTGG	TGCAACCTGATCTCCAGCCTA
Gd2_8=ChIP L19_17	CCATTCTTTCCTGCGGTCAG	CAGCAGGCACTGAAACATGC
Gd2_9=ChIP L19_19	GCCCATTTGGAATTGGCTCT	TGGCAAGTTCTGTGTTTGCAG
Gd2_10=ChIP L19_20	AGGCAGGGATGGGAAGATGT	TCTTGCCTTAGGAAGCAGTGA
Gd2_11=ChIP L19_21	TGGCAGTAAGTGACACCGCTATT	TGGCTCTGAGGAGAGAAAGTGG
PuroGFP	**Forward Primer**	**Reverse Primer**
PuroGFP_4= PuroC	TCTGGACCACGCCGGAGA	CCAGGAGGCCTTCCATCTGT
PuroGFP_5=GFPA	AAGGGCGAGGAGCTGTTCA	CCGGTGGTGCAGATGAACTT
PuroGFP_6=GFPD	TCAAGATCCGCCACAACATC	ATGTGATCGCGCTTCTCGTT
ArtCGI	**Forward Primer**	**Reverse Primer**
ArtCGI_5=ArtCGI 9	CAGGAGTTGACTCGGCAGGT	CCAGATGCGGTTCAGGTGAC
ArtCGI_6=ArtCGI 10	GCGTTCCCTCCTACCCTAGC	CCTGGTTGGCAGTCGGTTAC
ArtCGI_7=ArtCGI 12	GAAGCCCACCAGGGTGTCTC	GGAGGCGGTCCAACTAGTCA
High CpG / Low G+C	**Forward Primer**	**Reverse Primer**
Hi/Lo 5= HighCpG 1a	CGATAGCGAGGTTGGCACTC	CGTCCGCATAATCTTCTGAACG
Hi/Lo 6 = HighCpG 4	AAGTTTGTCGCGCATGAAAGA	TCTGCGACAATCACGTTTGTTT
Hi/Lo 7= HighCpG 5	TGTTCGTGTACGCTTCGATCAT	CGATCTGAGAGGTCGCATCC
Low CpG / High G+C	**Forward Primer**	**Reverse Primer**
Lo/Hi 5= LowCpG 1	AAGCTGGACTTCCGTCATGC	GGCAGGCTCAACATGTCCTT
Lo/Hi 6= LowCpG 5	GGCTGGGTAGAGTCCCAAGG	GGTCACCTGCCCTGGAGAG
Lo/Hi 7= LowCpG 7	CCACCAGCATATGGCCTCTC	GGCCGGCATCCTACAGAAG

Table 2 ChIP primers for Q-PCR

ArtCGI	**Forward Primer**	**Reverse Primer**
BS_ArtCGI_6	TTGGTATTTATYGGGTGGGTATA	ATCCAACTAATCAAAAAT
High CpG / Low G+C		
BS_High_6	TGTGAAATTYGGTTTTATTTTTT	RTTTATTTTATCATTTATCAT
IAP elements		
IAP LTR	TTGATAGTTGTGTTTTAAGTGGT AAATAAA	AAAACACCACAAACCAAAATCTTCTA C

Table 3 Bisulfite primers

	Forward	**Reverse**
E19_2F (2)	CACTGACCCAGAGTGAACCA	TGGCAGGGAAGGTAGCTATG
Chr18 E19 Scr 3'HR 1F (3)	GCATCGCAAGAGGAAATGTG	GCTGCACTATTGCTTGTTGC
Col PCR PuroGFP F(4)	CTGCAAGAACTCTTCCTCAC	CTCGTAGAAGGGGAGGTTG

Chr18 E19 Scr 5'HR 1 (5)	CCCAGGAAATATTTATGGAATG	AGTTCTTGCAGCTCGGTGAC
E19 4F (6)	CCATCTTGCCAGTTCCTCAT	ACCCACCCCCTCAAAATAAC
pBeloBAC 11 F (7)	ACAGATTTGAGGGTGGTTCG	CGCTGCTTCACCTATTCTCC
Probe L19 5	ACAGGGAAGGGTCACTGTGT	TGGCACCAGTAACAGCTGAA
Probe L19 3	TTCAAAGAAAAGCTGTTTCCAA	TACTGCCAGCAATGGGAGAT
PuroGFP 1F+R	GTCACCGAGCTGCAAGAACTC	TTACTTGTACAGCTCGTCCATGC
ArtCGI 1	CTAGCACAGGAGTTGACTCGG	TGCTCCGGGAACGTCAC
High CpG/Low G+C	TCTATAGCACACGGGCCAAC	AATGCTCTCTCGTTCATTCGT
Low CpG/High G+C	AGCTTGGCCCCTGGGGGA	CAGGGGTTCAGCTGGCAG
L19 3' HR	CCCATCCTCTTCAGAGATACCC	ACACCCACGACACTGATTCG
L19 5' HR	AATGAACAGGGCTTGGCTTC	ACAATCTGCAGGGCATCTCC

Table 4 Primers used to screen for the integration of BAC+ CGI-like sequence in mouse ES cell genome

No neo PuroGFP	TATATCATGGCCGACAAGCA	GCTGCACTATTGCTTGTTGC
No Neo High CpG/Low G+C	ACGAATGAACGAGAGAGCATT	GGGTATCTCTGAAGAGGATGGG
No Neo Low CpG/High G+C	CAGGGCCTCAGTGAGAATCAT	GGGTATCTCTGAAGAGGATGGG
Scr_noNeo_ArtCGI_1F	CGAAGGCAGCCTCCTACTG	GGGTATCTCTGAAGAGGATGGG

Table 5 Primers used to screen for excision of selection cassette

	Forward	**Reverse**
Probe PuroGFP	GACGTAAACGGCCACAAGTT	GAACTCCAGCAGGACCATGT
Probe Art CGI	TACGGTCCGACTTTTGCACT	ACGTCGATTGTGCGACCT
Probe High CpG Low G+C	CTCCGACCAAAAACCGAACA	GCGAAAACGCACAACTACCC
Probe Low CpG High G+C	CCCACCCTAGGGGATGCTAA	GAAGCCCCTACCTCCAATGC

Table 6 Primers used to amplify southern blot probes

2.1.5. Cell-lines used and created in this study

E14.Tg2a mouse ES cells were used throughout this study. Cfp1-GFP tagged mouse ES cells were obtained from Francis Stewart. The cell line was generated by BAC transgenesis of a *Cfp1*-GFP BAC construct. This BAC contains the whole *Cfp1* gene with all the regulatory elements and GFP fused to the last codon of *Cfp1* in order to create a C-terminal GFP tag. *Cfp1*-/- mouse ES cells were created in the laboratory of David Skalnik (Carlone & Skalnik, 2001). *Dnmt3a/3b* double knock out mouse ES cells (DKOs) were used as discribed (Okano *et al*, 1999).

	Names in this thesis	Clone names	Location in liquid N_2
Wt mES cells + Gene desert 1+ PuroGFP	Cell line 1	1D2_C3	Tower1/Tray3/row10/column9
Wt mES cells + Gene desert 2 + ArtCGI	Cell line 1 Cell line 2 Cell line 3	1C1_D11 1D6_C11 2D3_D6	Tower8/Tray1/row5/column5 Tower8/Tray1/row3/column9 Tower8/Tray1/row4/column6
Cfp1-GFP mES cells + Gene desert 2 + ArtCGI	Cell line 1 Cell line 2 Cell line 3	F7_1B1 G7_2C5 G10_1A2	Tower1/Tray11/row8/column8 Tower1/Tray11/row8/column6 Tower1/Tray11/row9/column3
Cfp1-/- mES cells + Gene desert 2 + ArtCGI	Cell line 1 Cell line 2 Cell line 3	A1_1C5 C2_1A4 C10_1B1	Tower8/Tray7/row8/column1 Tower8/Tray7/row8/column3 Tower8/Tray7/row8/column8
Wt mES cells + Gene desert 2 + Low CpG / High G+C	Cell line 1 Cell line 2 Cell line 3	A6_A8 B11_B7 C3_F10/E12	Tower8/Tray1/row7/column9 Tower8/Tray1/row8/column5 Tower8/Tray1/row10/column1 Tower8/Tray1/row9/column9
Wt mES cells + Gene desert 2 + High CpG / Low G+C	Cell line 1 Cell line 2 Cell line 3	A3_C1 A6_G3 A9_1D2	Tower8/Tray7/row4/column1 Tower8/Tray1/row7/column1 Tower8/Tray7/row9/column8
DKO + ArtCGI	Cell line 1 Cell line 2	C3_F4 C11_C10	Tower1/Tray11/row1/column5 Tower1/Tray11/row1/column3
DKO + High CpG / Low G+C	Cell line 1	# 41	Tower1/Tray11/row10/column1

Table 7 Cell-lines created in this study
Selection cassette has been excised in all these cell lines

2.2. Methods

2.2.1. Bacterial methods

Transformation of *E. coli*

For transformation either 1 ng of plasmid DNA or 10 µl were mixed with 75 µl of thawed competent cell suspension and incubated on ice for 20 min. The cells were then heat shocked at 42℃ for 1.5 min and then returned to ice for 2 min. Pre-warmed 0.5 ml of LB media was added to the cells, which were then incubated with shaking at 37℃ for 1 h. Cells were then transferred to LB agar or LB broth containing selective antibiotics and incubated overnight at 37℃.

Plasmid preparation

Plasmid DNA was purified from DH5α *E.coli* an endonuclease deficient (endA) and recombination deficient (recA) strain, which helps stable maintenance of the insert sequence,

using either a Quiagen miniprep kit, Quiagen midiprep or maxiprep kit depending on the quantity of DNA needed.

Recombineering

Bacteria containing the gene desert BAC of interest was plated on a LB plate containing chloramphenicol and incubated o/n at 37°C. The next day 2 or 3 single colonies were picked, inoculated in 15 ml tubes containing 1.0 ml LB medium with chloramphenicol and incubated at 37°C o/n with shaking. On day 2 a 2 ml ependorf tube containing fresh 1.4 ml SOC medium was set up and inoculated with 30 µl of fresh overnight culture. A hole was punctured in the lid for air. Bacteria were cultured for 2-3 h at 37°C in thermomixer, shaking at 1000 rpm. In the meantime cuvettes for electroporation and sterile water were chilled on ice and bench top centrifuge was cooled to 2°C. In order to electroporated the Red/ET plasmid into the cells containing the BAC cells were centrifuged for 30 sec at 11000 rpm, the supernatant was discarded and the pellet resuspended in 1ml of chilled H_2O. The centrifugation and resuspension step was repeated 2 times. After the last wash the pellet was resuspended in 40 µl of water and 20-40 ng of the Red/ET plasmid was added. As a control one reaction without Red plasmid was set up. Cells were electroporated at 1350 V, 10 µF, 600 Ohms using an electroporation cuvette with a slit of 1 mm. Cells were resuspended in 1ml of SOC and recovered for 1h at 37°C with shaking. Cells were divided on two plates (100µl and 900µl) containing chloramphenicol for selecting the BAC and tetracycline for selecting the Red plasmid. Plates were incubated at 30° for at least 15h protected from light.

On day 3 2-3 colonies were picked and cultured in LB+ chloramphenicol and tetracycline over night at 30°C protected from light. On day 4 a 2 ml ependorf tube containing fresh 1.4 ml SOC medium with the appropriate antibodies was set up and inoculated with 30 µl of fresh overnight culture. As a control one tube was set up with cells containing the BAC without the Red/ET plasmid. A hole was punctured in the lid for air. Bacteria were cultured for 2-3 h at 37°C in a thermomixer, shaking at 1000 rpm. In the meantime cuvettes for electroporation and sterile water were chilled on ice and bench top centrifuge was cooled to 2°C. Then 50 µl of 10% L-arabinose was added to one of the experimental tubes and to one of the control tubes in order to induce the expression of the Red/ET recombination proteins. Another tube was left without induction as negative controls. All samples were incubated at 37°C, shaking for 1 h. Afterwards cells were centrifuged for 30 sec at 11000 rpm, the supernatant was discarded and the pellet resuspended in 1ml of chilled H_2O. The centrifugation and resuspension step was repeated 2 times. After the last wash the pellet was

resuspended in 40 μl of water and 200-300 ng of linearized DNA containing the CGI-like sequence, a kanamycin selection cassette and a 3' and 5' homology arm for recombination was added. Cells were electroporated at 1350 V, 10 μF, 600 Ohms using an electroporation cuvette with a slit of 1 mm. Cells were resuspended in 1ml of SOC and recovered for 1-2h, during which time recombination occurred, at 37°C with shaking. Cells were divided on two plates (100μl and 900μl) containing chloramphenicol for selecting the BAC and kanamycin for selecting the CGI-like construct. Plates were incubated at 37° o/n. On day 6 colonies were picked and either screened by colony PCR for the successful recombination or another o/n culture was set up to do minipreps and control digests.

2.2.2. DNA manipulation

Synthesis of CGI-like sequences

CGI-like sequences were designed using the "Random Sequence by CpG parameters" tool found at the Wellcome Trust Centre for Cell Biology Galaxy server. Specifc paramaters for different CGI-like sequences were chosen regarding CpG frequency, G+C content and length. SP1 binding sites were avoided and restriction enzyme sites for *HindII* and *EcoRI* were attached to the 5' and 3' end respectively. The randomly obtained DNA sequence were synthesized by GeneArt ® Gene Synthesis from Life Technologies. The DNA was delivered as 5 μg lyophilized plasmid DNA.

DNA gel electrophoresis

DNA was resolved by gel electrophoresis using the Bio-Rad Sub-Gel system. Agarose gels (1-2%; depending on the size of the DNA fragment to be resolved) were prepared with TAE containing 0.5μg/ml ethidium bromide, a DNA intercalating agent that fluoresces in ultraviolet (UV) light. DNA samples were prepared in orange G loading buffer and size markers (1kb plus ladder; Thermo Scientific) in supplied 6x loading dye. Samples were loaded into the wells of the gel. Gels were run at constant voltage (80-110V) in TAE electrophoresis buffer and visualised under UV light.

Restriction enzyme digest

DNA digest were carried out as per manufacturers instructions (NEB). Briefly, DNA was diluted in appropriate digestion buffer, supplemented with 100 μg/ml Bovine Serum Albumin where appropriate and digested with 6U of restriction endonuclease per μg of DNA. Reactions were typically incubated at 37°C for 1-2hours.

Ligation

Standard DNA ligation reactions were performed in a total of 20 µl containing 100 ng linearized vector, insert DNA (3x molar excess), T4 DNA ligase buffer and 1 µl T4 DNA ligase (NEB) and incubated o/n at 16°C.

DNA extraction and precipitation

DNA was extracted through mixing with 1 volume of phenol:chloroform:isoamyl alcohol (IAA) before precipitation by either 1 part isopropanol at RT or 3 parts 96% ethanol (EtOH) at -80°C along with 1/10th volume of NaOAC. After centrifugation at 15,700g for 15 minutes pellets were washed with 70% EtOH before allowing to air-dry in a fume hood for 5 minutes. Once all residual ethanol has evaporated the DNA pellets were resuspended in 0.1M TE and stored at -20°C.

Gel Extraction

Gel extraction was used to purify a homogeneous population of DNA fragments for cloning or probe preparation. Fragments were resolved by agarose gel electrophoresis, cut out and extracted using the Zymoclean Gel DNA recovery kit.

Measurement of DNA concentration

DNA solutions were measured at OD260nm and OD280nm using a Nanodrop-1000 spectrophotometer (Thermo Scientific). The purity of the DNA was determined using OD260nm:OD280nm ratio, with a value greater than 1.8 indicating the absence of protein or phenol contaminants from the sample. An automated read out of the DNA concentration was determined through the use of Beer's law [Concentration in ng/µl = (Absorbance OD260nm x Extinction coefficient dsDNA 50ng/µl/cm) / pathlength cm].

Polymerase Chain reaction (PCR)

PCR was used to amplify DNA molecules of interest from a starting template and was used primarily to test primers for qPCR, for Colony PRC to screen for successful recombination or BAC integration and in bisulfite genomic sequencing. Reactions were generally carried out in a 20µl volume and consisted of DNA template (0.1-50ng), 400nM forward and reverse primers, 400µM dNTPs (Abgene), 2.5mM MgCl2, Red Hot Taq reaction buffer (Abgene) and 1.5U Red Hot Taq (Abgene). PCR amplification of bisulfite-treated DNA was carried out in the presence of 3% dimethyl sulfoxide (DMSO). PCR reactions were set up on ice. PCR cycling was carried out on a G-Storm thermal cycler and typical conditions were as follows: initial denaturation at 94°C for 2min followed by 30 cycles of denaturation at 94°C

for 30sec, primer annealing at appropriate temperarure for 30sec and primer extension at 72°C for 30sec. An additional 72°C primer extension phase for 5min was carried out at the end of the 30 cycles to amplify any incomplete DNA molecules. In general, to determine the optimal annealing temperature (Tanneal) for a particular primer pair a range of temperatures was tested (using the gradient feature of the thermal cycler over a range of 53-64°C). PCR amplification of bisulfite-treated DNA was carried out for 40 cycles with longer step times: 94°C for 40sec, Tanneal for 50sec and 72°C for 50sec. PCR products were resolved by agarose gel electrophoresis.

DNA sequencing

DNA sequencing was performed using the BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems). Sequencing reactions were carried out in a volume of 10µl and contained 4µl of 2.5x sequencing buffer, 2µl BigDye Terminator, 500nM primer and 3.5µl DNA (prepared for sequencing as above). Reaction conditions consisted of initial DNA denaturation at 94°C for 10sec followed by 24 cycles of DNA denaturation at 94°C for 30sec, primer annealing at 50°C for 20sec and extension at 60°C for 4min. Sequencing reactions were then passed onto the Gene Pool Sequencing Service (School of Biological Sciences) where they were cleaned up and run on an ABI 3730 capillary sequencer.

Preparation of BAC DNA for transfection of mouse ES cells

BAC DNA was prepared using the Nucleobond® Kit from Machery Nagel following the manufacturers instructions. BAC DNA was linearized using an appropriate restriction enzyme. An aliquot of the linearized BAC DNA was run on an agarose gel to check for complete linearization and the rest was purified using phenol:chloroform:IAA extraction and ethanol precipitation.

Genomic DNA extraction from ES cells

Cells were trypsinized, washed in PBS and pellet resuspended in 1ml lysis buffer containing 100 mM NaCl, 10 mM TrisCl, pH 8, 25 mM EDTA, pH 8, 0.5% SDS and 0.1 mg/ml proteinase K added freshly. Cells were incubated o/n at 55°C. The next day RNAse A was added and incubated for 1h at 37°. Samples were extracted twice with an equal volume of phenol/chloroform/isoamyl alcohol. The aqueous phase was transferred to new tube and DNA precipitated with 1/10 volume of 3 M NaAc and 1.5 volume of 100% ethanol. DNA was washed with 70% ethanol and resuspended in an appropriate volume of TE.

Preparation of genomic DNA from 96-well plates

Once a colony growing in a 96-well has reached confluency, the medium was aspirated off and 100 µl of lysis buffer were added and returned to the incubator over night. Cell lysates were stored at 4°C until all cells were lysed. Then 150 µl of ice-cold isopropanol were added, the plate was sealed, inverted 2-3 times and placed on a shaker. Plate was spun down for 30 min, isopropanol poured off and 50 µl 70% Ethanol was added and plate centrifuged for 30 min. Supernatant was removed and plate placed on a heat block for 3 min to evaporate excess alcohol. Finally 50 µl of TE was added and placed over night at 4°C to promote resuspension.

Determining copy number by qPCR

In order to determine the copy number of BAC+CGI-like sequence integrated in the mouse genome a standard curve of BAC plasmid DNA was created, where the BAC is present at 300,000 copies, 30,000 copies, 3,000 copies, 300 copies and 30 copies. The mass of a single plasmid was calculated multiplying the plasmid length with the mass of a double stranded DNA molecule. This number was used to calculate the required amount of plasmid DNA needed to achieve the different copy numbers. The Ct values obtained from the qPCR were used to blot a standard curve. This standard curve was used to calculate the copy number of the clones of interest per cell. Of each clone 200ng DNA was used per qPCR reaction and it was assumed that each cell contains 6pg of DNA. As a control and to normalize to a known copy number the Sox2 gene, a gene that exists as a single copy per haploid genome (or 2 copies per human cell), was used.

To make a standard curve for Sox2 mouse genomic ES cell DNA in which the gene of interest is present at 300,000 copies, 30,000 copies, 3,000 copies, 300 copies and 30 copies was prepared. First the mass of DNA per mouse genome was calculated. Then the mass of the genome was divided by the copy number of the gene of interest per haploid genome. With this value the mass of gDNA containing the 300,000 to 30 copies was determined and the dilution serial pipetted. For the qPCR 5µl of gDNA was used per reaction and the obtained standard curve was used to calculate the amount of Sox2 present in the clones measured. This value was used to normalize the obtained values for the BAC integrations as it is assumed that Sox2 is present in 2 copies per cell.

Southern Blot

Between 20 and 25 μg of DNA was digested o/n at 37°C with the appropriate restriction enzyme in a volume of 30-50 μl. The next day samples were run on a 0.7% gel in 1x TAE during the day at around 50V. Hyb+ buffer was prepared and frozen in aliquots. Gel was photographed with a fluorescent ruler, excess gel was cut and remaining gel put into HCl solution for 15 min while shaking. Gel was rinsed with water and neutralized in 0.4M NaOH for 45 min. In the mean time filter paper and membrane were cut to size of gel and membrane was pre-wet and incubated in NaOH for 5 min. A dry plot was assembled and left o/n at RT. The next day 25 ml of Hyb+buffer was pre-warmed to 65°C. Filter was placed in 2x SSC for 5 min, pre hybridized for 2h at 65°C. Around 50 ng of probe was labelled using the megaprime kit. Unincorporated nucleotides were removed using G50 column. Labelled probe was added to 15 ml of hyb+ buffer and leave rotating at 65° C o/n. The next day blot was washed twice with pre-warmed 2xSSC. Level of radioactivity was checked with Geiger counter and if still very hot, blot was washed with 1x SSCe washing once more for 30 min. Membrane was put into phospho imager and exposed for 1h–over night.

Quantitative PCR (qPCR)

qPCR was used to assess enrichment of specific regions in ChIP samples and to determine the copy number of BAC DNA into the mouse genome. To carry out qPCR SYBR Green technology was used. SYBR Green is a fluorescent dye that binds to DNA and, when included in a PCR reaction, allows fluorescence to be used as a read-out for the amount of DNA synthesised in real-time. qPCR reactions (10μl) contained SYBR Green SensiMix (Quantace), 250nM primers and 3ul ChIP DNA or 5ul DNA for the determination of copy numbers. PCR was carried out using a Roche Lightcycler and cycling conditions were as follows; initial denaturation at 94°C for 10min followed by 45 cycles of denaturation at 94°C for 10sec, primer annealing at Tanneal for 10sec and primer extension at 72°C for 15sec. Data were collected at the end of each amplification cycle. After amplification, melting curves for PCR products were generated by denaturing the DNA at 94°C for 1min and then increasing the temperature from 35°C to 95°C in increments of 0.1°C. A single melt curve indicates the presence of a single PCR product whilst multiple melt curves could indicate the presence of multiple PCR products or primer dimers that would interfere with DNA quantification. Using the Roche Lightcycler software, SYBR Green fluorescence measurements were plotted relative to cycle number to generate curves for each PCR reaction and the 2nd derivative maximum method was used to determine the cycle threshold values (Ct) for each sample. An arbitrary measure of DNA quantity was calculated using the

formula 2^{-Ct}. This value was then compared to that of a reference sample (for example input DNA in a ChIP reaction). qPCR reactions were carried out in duplicate or triplicate.

Bisulfite genomic sequencing

Bisulfite conversion of genomic DNA was carried out using the EpiTect Bisulfite Kit from Quiagen. The converted DNA was used for PCR amplification of region of interest and reaction run on an agarose gel. Band was excised and used for cloning using the Stratage blunt end cloning kit. The next day colonies were screened for the integration of the fragment and positive clones were sent for sequencing

2.2.3. Protein manipulation

Chromatin Immunoprecipitation (ChIP)

Chemical crosslink of chromatin was performed *via* addition of 1% formaldehyde for 10 minutes at 37°C directly to the medium in tissue culture dishes. Then, 500µl 2.5M glycine was added in order to stop the crosslinking process. After 5 minutes incubation while shaking, cells were washed twice with 1x PBS. If acetylation was examined, sodium butyrate, an HDAC inhibitor was added. In the next step, cells were centrifuged for 5 minutes at 1.200 rpm at 4°C and washed in wash buffer A and B (each buffer contained PMSF to a final concentration of 1µg/ml, 1x Proteinase Inhibitor Complete Mix (Roche) to inhibit protease activity and 10mM sodium butyrate). After each wash, cells were kept on ice for 10 minutes. Finally, cells were resuspended in lysis buffer, also containing PMSF, Proteinase Inhibitor Complete Mix and sodium butyrate at concentrations indicated above. In the next step, chromatin was sheared to an appropriate size (400-600 base pair fragments) by sonication of the cell lysate using a twin Bioruptor for 15 cylces, 30 sec on, 30 sec off on high setting. After completion of sonication lysate was centrifuged for 10 minutes at 14.000 rpm at 4°C to collect cellular debris. In order to check for successful sonication 10 µl of sonicated chromatin were used and 100 µl Chelex 100 Resin by BiorRadb (10%; 0.1g in 1ml H2O) were added. Mixture was vortexed, boiled for 10 min at 99 °C, vortexed and cooled down at RT for 2-3 min. Then 1µl RNAse was added for 30 min at 37 °C and then 2µl Proteinase K for 30 min-1h at 55 °C. Samples were spun down and 50 µl supernatant was loaded with loading dye on a 2% agarose gel and let run shortly. Chromatin concentration was measured by Nanodrop and 30 µg were used for ChIP with antibodies against histone modifications. For proteins 150 µg chromatin was used. For the input fraction, 15 µg was set aside, lysis buffer was added to 50 µl and input chromatin was stored at 4 °C. For the ChIP

required amount of chromatin was made up to 900µl with dilution buffer in the presence of PMSF, Proteinase Inhibitor Complete Mix and sodium butyrate. Antibodies were added as specified and incubated o/n rotating at 4 °C. Per IP 50 µl of magnetic protein G Life Technology dynabeads were used. Beads were washed on a magnetic rack 3 x with TE, blocked with 1/10 volume of BSA and left rotating o/n at 4 °C. The next day 50µl of blocked beads were added to the supernatant of IPs and incubated rotating for 1-4 hours at 4 °C. IPs were washed 3x with 1ml ice-cold wash buffer 1, 2x with wash buffer 2 and 1x with TE. After each wash IPs were spun for 2 min at 1200rpm, the supernatant was removed, the next wash buffer added and incubated rotating at RT for 2 min. On last wash, all supernatant was removed and 100µl of freshly prepared 10% Chelex was added. To the input fraction 50 µl of freshly prepared 10% Chelex was added. Samples were boiled at 100°C for 12 min, left to cool at RT 2 µl of 20mg/ml Proteinase K was added to each sample, incubated at 55 °C for 30' at 1000 rpm on thermomixer. The samples were boiled at 100 °C for 10'. Tubes were shortly spun and 60 µl supernatant was transferred to fresh tube. Each sample was made up to 300µl total volume with 10mM Tris pH8+0.1 mM EDTA, stored at -20 °C and 3 µl DNA was used for Q-PCR. Input was diluted 1:10 and also 3 µl were used for Q-PCR.

2.2.4. Manipulation of mouse ES cells

Culturing mouse ES cells

E14.Tg2a mouse ES cells were cultured in Glasgow MEM medium supplemented with 15% FBS, 1% sodium pyruvate, 1% non-essential amino acids, 0.1% β-mercaptoethanol, 100U/ml penicillin, 100 µg/ml streptomycin and leukemia inhibitory factor (LIF; gift from J. Guy) on gelatinised tissue culture plastic ware. Cells were split when reaching around 80% confluency in a ratio between 1:8-1:16 using trypsin. Cells were cultured at 37°C with 5% CO_2.

Transfection mouse ES cells with BAC DNA

Between 0.5 and 2 µg of linearized BAC DNA was used to transfect 60% confluent mouse ES cells growing in a 6-well plate using Lipofectamine™ LTX Plus™ by Invitrogen. In short, DNA was made up with OptiMem (from GIBCO) to 500 µl, 2.5 µl PLUS reagent was added and incubated for 5 min at RT. Afterwards 6.25 µl Lipofectamine was added and incubated for 30 min at RT. The solution was then added drop-wise to the ES cells. Cells were split in a range of different ratios (25-0.1% of transfected cells) 24 h after transfection and plated onto 10 cm^2 dishes. The next day selection medium containing the appropriate

antibody (G418 250 μg/ml or Blasticidin 3 μg/ml) was added to the cell. Cells were grown until colonies were ready to be picked (8-10 days post transfection).

Colony picking

Colonies were picked by sucking them into a tip containing 5 μl of PBS using a p20 pipette. Each colony was transferred into one well of a v-bottom 96-well plate. Once 48 colonies were picked, 20 μl of trypsin was added to each well and cells were trypsinized for 5 min at 37°C and dissociated by pipetting up and down to obtain single cell suspension. Cells were transferred to a new flat-bottom 96-well plate and cultivated in selection ES cell medium until most wells reached confluency.

Splitting and Freezing of 96-well plates

Colonies growing in a 96-well plate were split by transferring half of the cells in a new 96-well plate. The other half was transferred into 24-well plates. Cells in the 96-well plate were grown to confluency and used for genomic DNA extraction. Cells in the 24-well plates were grown to confluency, trypsinized, resuspended in 100 μl of normal ES cell medium, transferred to new 96-well plate, 100 μl of freezing medium (80% FCS, 20% DMSO) were added, plate was sealed and stored at -80°C.

Excision of the selection cassette by Flp/Dre

Cells for excision of the selection cassette were grown to confluency in a T75 flask without antibiotics, fed 4 h before electroporation, trypsinized, washed in PBS, centrifuged for 5 min at 1300 rpm and resuspended in 700 μl PBS. Then 50 μg of circular plasmid containing Flp or Dre were added, mixture was transferred to a cuvette and electroporated at 250V and 500 μF using a BioRad electroporator. Cells were left to recover for 20 min at RT and diluted in 10 ml of medium. Electroporated cells were transferred in different dilutions. For a typical experiment 50%, 25%, 10%, 5%, 1%, 0.1% and 0.01% of the electroporated cells were transferred to 10-cm dishes. The next day medium was changed and cells cultured until colonies were big enough for picking.

As an improved version, a transient selection step with puromycin 0.8 μg/ml introduced. After the cells were electroporated, medium was changed the next day and 24h later puromycin was added for 48h after which cells were grown in normal medium until colonies were ready to be picked. For the DKO cell-lines this step was omitted, as the cells were already resistant to puromycin.

Differentiation of mES cells into neural precursor cells

Rapidly growing mES cells with good ES cell morphology were plated in bacterial dishes (4x106 cells/dish) in 15ml EB medium (normal ES cell medium with only 10% FBS and no LIF to encourage the formation of embryonic bodies). After 4 days in EB medium RA was added to begin neuronal differentiation. Medium was changed every 2 days. On day 8 cells were disrupted, trypsinized and used for X-linking.

3. Does an artificial CGI impose an alternate chromatin structure in a gene desert?

3.1. Introduction

Despite many years of intensive research the functional significance of CpG islands (CGIs) with respect to chromatin structure and transcription is still not clear. This thesis will therefore aim to analyse the role of the dinucleotide CpG as a signalling module.

In a recent study Thomson and co-workers introduced an artificial, promoterless DNA sequence that shows typical characteristics of CGIs regarding length, CG content and CpG frequency into the genome at sites that lack H3K4me3. The DNA inserted at the untranslated region of the *Nanog* gene comprises the promoterless EGFP (700bp with 60 CpGs) coding sequence adjacent to a puromycin resistance gene (600bp with 93 CpGs). We term this insertion "PuroGFP" and the cell line "*Nanog*-PuroGFP". In another experiment the promoterless eGFP coding sequence was fused to the 5' end of the X-linked gene *MeCP2*, called "*MeCP2*-eGFP". These CpG island-like insertions recruited Cfp1 and created new marks of H3K4me3 despite the absence of a promoter. Notably no RNA Polymerase was detected (Thomson *et al*, 2010). This data indicates that one function of non-methylated CpG islands is to genetically influence the local chromatin environment by interaction with Cfp1 and possibly other CpG-binding proteins (see Figure 2.2.4-1 taken from (Thomson *et al*, 2010)).

A caveat of this study is that the CGI-like sequences were in both cases introduced into a transcriptional unit. Therefore it cannot be excluded that the observed effects on Cfp1 recruitment and chromatin establishment are due to its integration site and not a general feature of CGI-like sequences. In order to address this issue the present work focuses on the analysis of different CGI-like sequences in so-called gene desert regions, which are characterized by the absence of CpG islands and modified histones and which do not have any indication of transcription. Additionally, the relative contribution of CpG frequency versus G+C content will be investigated.

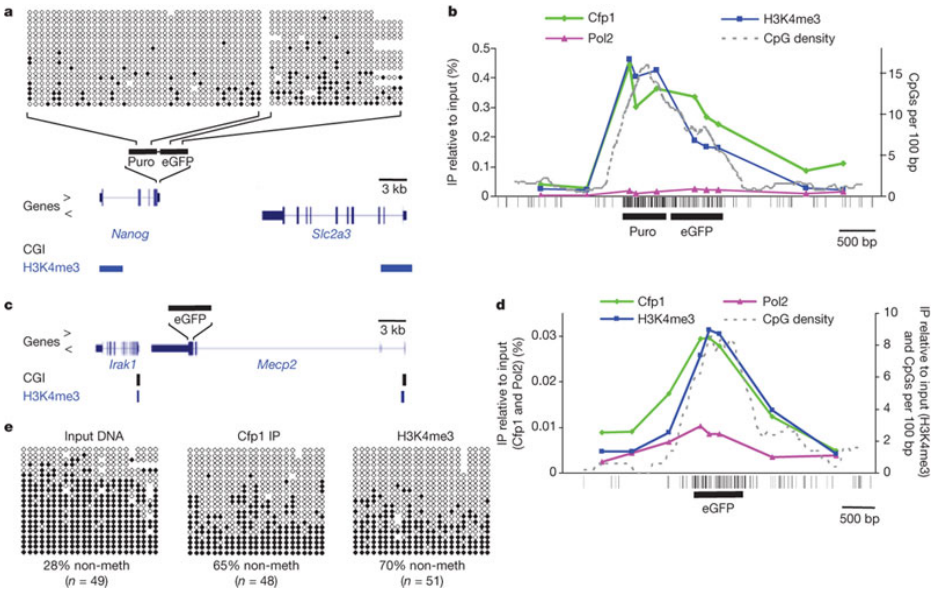


Figure 2.2.4-1 Artificial promoterless CpG-rich sequences recruit Cfp1 and generate new H3K4me3 peaks in mouse ES cells.
a: The ES cell line carries adjacent promoterless eGFP and bacterial puromycin-resistance sequences (black bars) inserted together within the untranslated region of the *Nanog* gene. c: The eGFP sequence was inserted within the untranslated region of the *Mecp2* gene. The positions of CGIs and H3K4me3 peaks at this locus in wild-type ES cells are shown below the map. DNA methylation within the insertion was determined by bisulphite sequencing of 306-bp (eGFP) and 275-bp (Puro) segments of the inserted sequence. b/d: ChIP qPCR across the region containing the insertion using antibodies against Cfp1, RNA polymerase II (Pol2) and H3K4me3. The dotted line plots CpG density in a 500-bp window with a 100-bp step size. Vertical strokes below the graph indicate positions of CpGs. e: Bisulphite sequence analysis determined the methylation status of input and DNA immunoprecipitated by the Cfp1 and H3K4me3 antibodies. The percentage of non-methylated CpGs is shown below each panel (taken from (Thomson *et al*, 2010)).

3.2. Results

3.2.1. A novel mark of H3K27me3 is created at a promoter-less CGI-like sequence

Previously the influence of a promoterless CGI-like sequence on H3K4me3 establishment alone was analysed (Thomson *et al*, 2010). As it was recently published that GC rich sequences are attracting Polycomb proteins (Mendenhall *et al*, 2010) we wondered if the aforementioned *Mecp2*-GFP and *Nanog*-PuroGFP tagged cell-lines are not only establishing a novel peak of H3K4me3 peak but also H3K27me3. Firstly, we wanted to confirm that a promoterless CGI-like sequence creates a novel peak of H3K4me3 over the inserted sequence in the recombinant cell lines used in the original study. Indeed, as Figure 3.2.1-1 shows both cell lines display H3K4me3 over the insertion. Several control regions were included in each experiment in order to provide information about the level of enrichment

observed over the integrated sequence in relation to active and bivalent gene promoters. Promoter regions of highly expressed house keeping genes such as *ActinB* or *GAPDH* as well as the pluripotency gene *Sox2* were used as examples for genes that display high levels of active chromatin marks and RNA Pol II at their transcription start sites. Two genes from the *Hox* cluster, *HoxC8* and *HoxA9*, were used as examples for bivalent genes, which show both H3K4me3 and H3K27me3 marks at their promoter regions. An inconspicuous intergenic region on mouse chromosome 15 (denoted as "m15") that does not exhibit any signs of transcription or H3K4 and H3K27 methylation was used as a negative control region (note that not all controls have been used in each experiment) (Clouaire *et al*, 2012). The H3K4me3 levels observed at the *Mecp2*-eGFP and *Nanog*-PuroGFP insertion were slightly higher than at the HoxC8 bivalent gene promoter HoxC8. As expected H3K4me3 levels over the active gene promoters *Sox2* and *ActinB* are much higher, whereas the negative intergenic control region on mouse chromosome 15 (m15) showed only very low levels of H3K4me3. When looking at the H3K27 modification, a novel mark of H3K27me3 can be observed over the inserted DNA fragments in both cell lines. Consistently, Suz12, a member of the PRC2 complex, that is responsible for establishment of H3K27me3, was also found over the inserted CGI-like sequences. In both cell lines H3K27me3 and Suz12 levels were lower over the inserted CGI than over the *HoxC8* bivalent gene promoter, but still clearly higher than adjacent flanking regions.

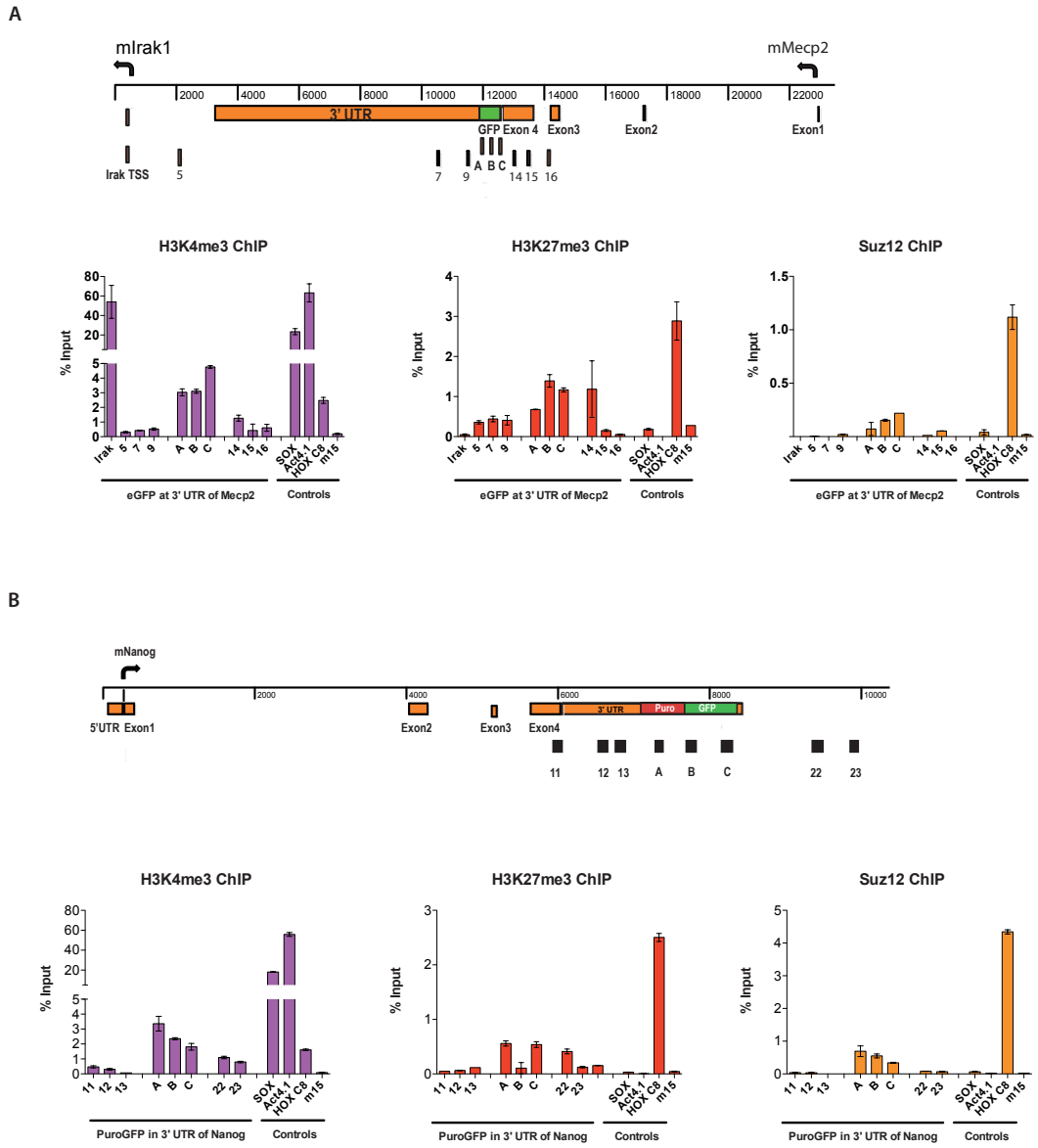


Figure 3.2.1-1 H3K4me3 and H3K27me3 marks are present at Mecp2-eGFP and Nanog-PuroGFP

Panel shows gene structure of Mecp2 (A) and Nanog (B) and depict the inserted CGI like sequence (GFP and PuroGFP) in 3'UTR of respective gene. Arrows indicate TSS and direction of transcription. Black boxes indicate position of primers used in ChIP qPCR. ChIPs with antibodies against H3K4me3, H3K27me3 and Suz12 were performed. Y-axis: % of Input. Controls: TSS of active genes Sox2 and Act4, of bivalent gene HoxC8; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates.

3.2.2. Insertion of the PuroGFP artificial CGI in gene desert by Recombination mediated cassette exchange

We wanted to test if the effect of a promoterless CGI-like sequence on chromatin structure was due to its integration site within genes. We therefore introduced the same PuroGFP

sequence as previously into a gene desert region and asked if a novel peaks of H3K4me3 and H3K27me3 were nevertheless created. Only the CGI-like sequence PuroGFP was chosen as this construct stayed unmethylated in contrast to the eGFP construct, which gained dense DNA methylation to about 70% (Thomson *et al*, 2010).

The initial approach to create a mouse ES cell line containing the puroGFP sequence in a gene desert was to employ recombination mediated cassette exchange (RMCE) (see Figure 3.2.2-1). This method allows targeted introduction of different DNA sequences into the same genomic locus. It is based on the use of two different targeting constructs where the first is introduced into the genome by homologous recombination. This creates a platform in a defined genomic locus where different CGI-like sequences can be introduced via a cassette exchange mediated by Cre recombinase. In this manner only one homologous recombination event is required for inserting several DNA sequences at the same locus. The first targeting construct contains a neomycin resistance cassette flanked by heterotypic Lox sites for selection in G418 media. The advantage of heterotypic Lox sites is that LoxP and Lox511 do not recombine with each other but rather they allow a cassette exchange. Furthermore it possesses the 5' part of the hypoxanthine-guanine phosporibosyltransferase (hprt) gene that is non functional in this construct but can be activated by the second construct. It consists of the remaining 3' region of the *hprt* gene and a multiple cloning site for the introduction of the CpG islands between heterotypic lox sites. This construct will be then used for transfecting the mES cells that contain the docking platform at the desired locus. A plasmid that expresses the recombinase cre will be co-transfected with the construct containing the CpG island. This will result in a cassette exchange between the first and second construct at the heterotypic lox sites. Cells that have undergone a successful cassette exchange will be no longer resistance to G418 as the neomycin resistance gene is not present anymore; instead the *hprt* gene is reconstituted. Cells with a functional *hprt* gene are able to synthesise nucleic acids through a purine salvage pathway. As a result these cells will survive in hypoxanthine aminopterin tymidine (HAT) containing media, which blocks *de novo* nucleic acid synthesis and forces cells to use the alternate salvage pathway. Therefore cells can be double selected for HAT resistance and through replica plating as well for neomycin sensitivity. Once cells have been identified that contain the PuroGFP, FLP recombinase is used to excise the promoter of the *hprt* gene. This is necessary as the aim of this experiment is to test the effect of CpG rich sequences on the local chromatin environment in absence of promoters.

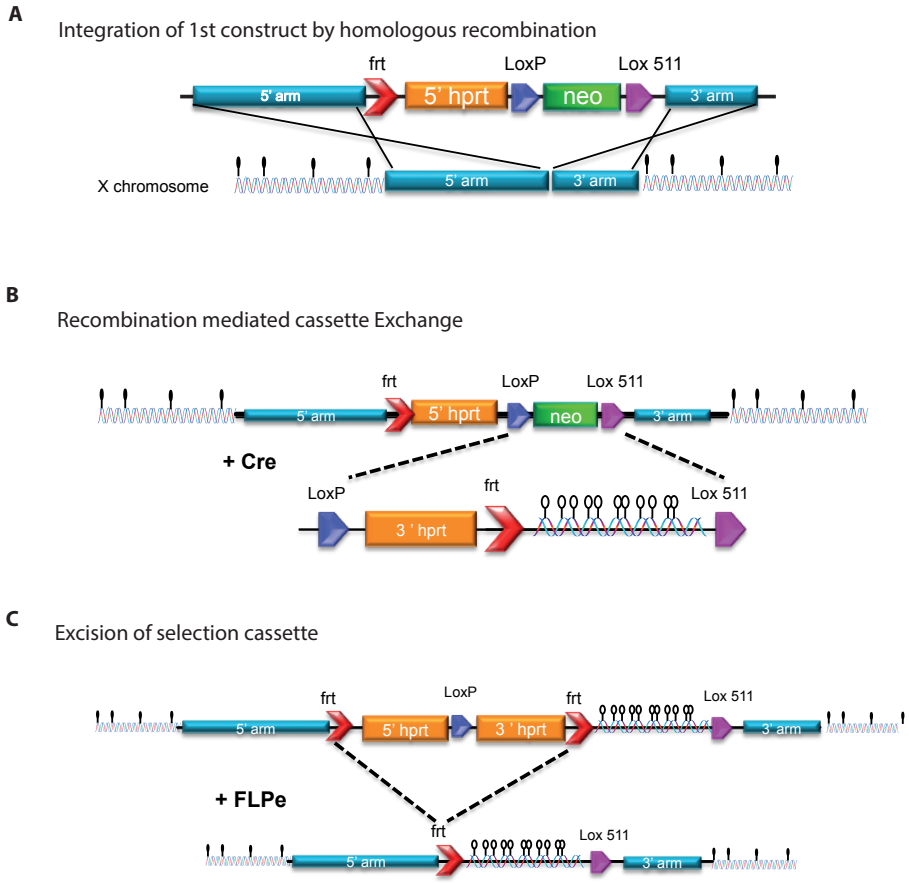


Figure 3.2.2-1 Overview of recombination mediated cassette exchange
A: The first construct is inserted into a defined location on mouse chromosome X creating a docking platform. B: Upon transfection with Cre cassette exchange takes place inserting a CGI like sequence and reconstituting a functional hprt gene. C: Upon transfection with Flp the hprt gene is excised leaving the CGI behind flanked only by frt and Lox511 sites.

We selected 2 different gene loci on the mouse X chromosome (X:44163816-44168855 and X:106815842-106823676) with no CpG island or H3K4me3 present. We chose the X chromosome since it is present as a single copy in male mES cells. In order to identify a region of the mouse genome that qualifies as a gene desert both data sets generated in our lab and published data from databases such as Ensemble or USCS were mined. CxxC affinity purification, a method that was developed in the lab to enrich for unmethylated CpG islands with subsequent high throughput sequencing allowed to identify regions within the genome with no CpG islands (Illingworth *et al*, 2008). ChIP of H3K4me3 and the initiation form of RNA polymerase II followed by Solexa sequencing permitted the identification of regions that lack open chromatin and show no sign of transcription According to the "Ensemble Mouse" assembly from 2007 there are no genes, pseudogenes or other transcripts annotated

within 200kb up- and downstream of the selected loci. Care was taken to avoid highly repetitive regions as this could impair homologous recombination. Homologous arms for Gene desert 1 and 2 were PCR amplified from mouse ES cells DNA and sequentially cloned into the first targeting construct. One homology arm was designed shorter, 1,7 and 1,3 kb respectively, in order to allow for PCR screening. The longer homologous arm was designed to have at least 3 kb. ES cells were transfected with the 1st construct and 500 colonies were picked to identify clones that have undergone homologous recombination by PCR and Southern blot. However such a clone could not be identified.

One potential reason for the failure to obtain cells that have undergone homologous recombination is that introduction of the docking platform into a gene desert region might be less frequent than normally expected. According to the definition we tried to identify regions that are devoid of genes and histone modifications that are associated with open chromatin. The regions chosen therefore are rich in heterochromatic marks, which presumably strongly impaired the frequency of homologous recombination. Additionally the length of the homology arms might have been too short to allow efficient homologous recombination. As a consequence we decided to employ a different strategy to obtain a mouse ES cells line that contains a CGI-like sequence in a gene desert.

3.2.3. Insertion of puroGFP into gene desert by random integration into the mouse genome

In order to circumvent the problem of inefficient homologous recombination in a gene desert an approach was chosen that relies on the random integration of a gene desert containing the puroGFP sequence into mouse ES cells. This method is based on first introducing the CGI-like sequence into a human gene desert containing BAC by recombineering in bacteria. Subsequently the linearized human gene desert containing the puroGFP sequence is randomly integrated into mouse ES cells. After selection for integration of the construct the selection cassette can be excised using Flp-recombinase. Finally the chromatin state of different cell lines containing only the CGI with a remaining frt and loxP site in a gene desert can be analysed by ChIP. Figure 3.2.3-1shows an overview of this method.

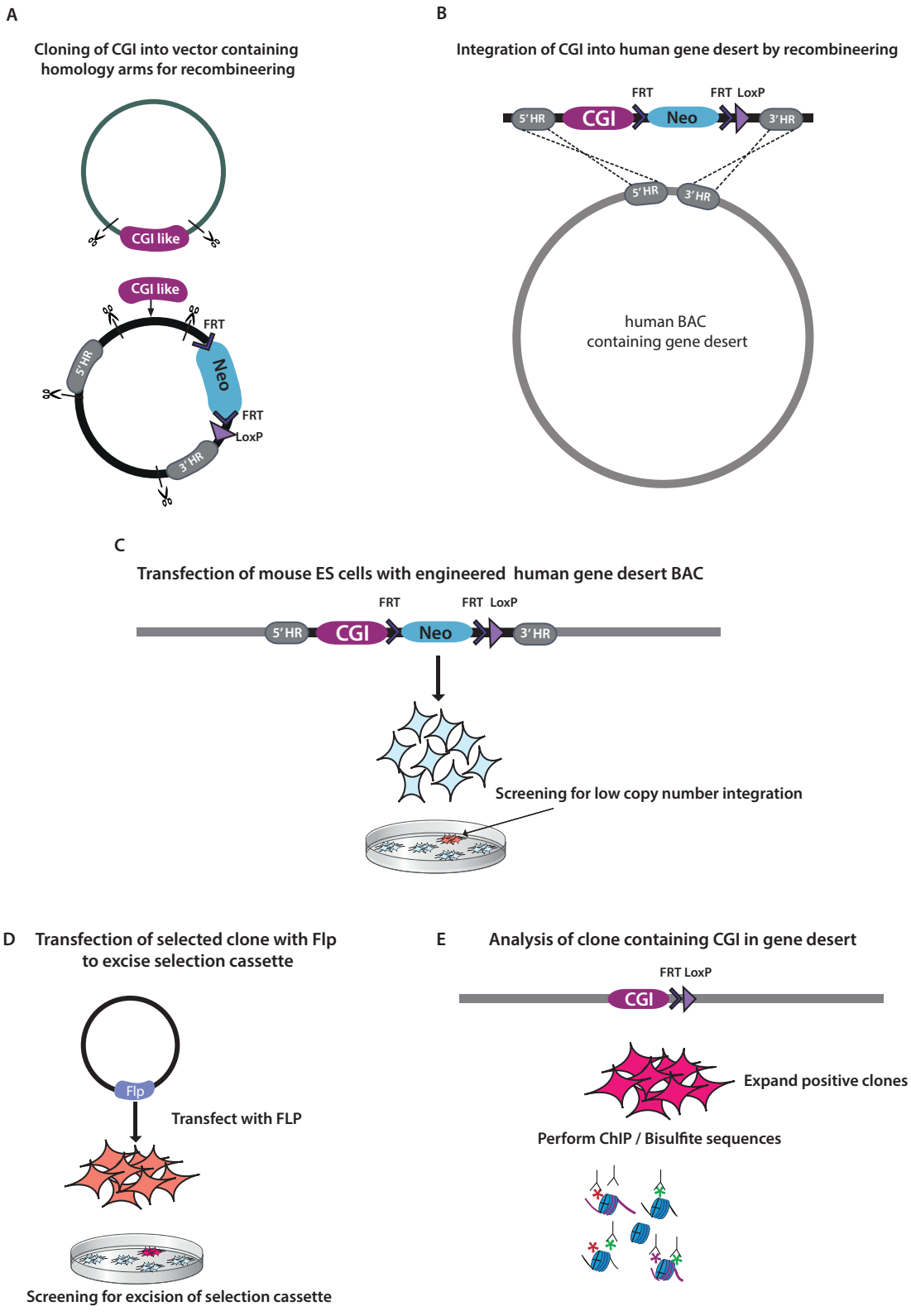


Figure 3.2.3-1 Insertion of CGI in gene desert into mouse genome by random integration

A: Cloning of CGIs into plasmid into which the homology arms for the gene desert and a selection cassette have been cloned. B: The linearized construct from A is used to transfect a human BAC containing a gene desert. Integration occurs via homologous recombination in bacteria (recombineering).

C: The linearized construct from C is used for transfecting mouse ES cells. Colonies with a random integration of the BAC are screened for low copy integration. D: Positive clones are transfected with Flp/Dre to excise selection cassette. Colonies are screened for successful excision by PCR and Southern blotting. E: Clones containing a CGI in gene desert without selection cassette are used for analysis of chromatin modification and DNA methylation. The gene desert regions either side of the CGI-like sequence provides insulator function that protect against positional effects and allow to study the influences of base composition on local chromatin establishment without stimuli from adjacent genes.

3.2.3.1. Identification of human gene desert

In a first step a human gene desert was identified to ensure unambiguous discrimination between the endogenous mouse genome and the newly integrated DNA. Genome wide H3K4me3 and RNA PolII ChIP-seq data of human ES cells and data from CAP of human sperm generated in our laboratory was used to identify a 132.46 kb region on chromosome 18 that satisfied the criteria of a gene desert and that we termed "Gene desert 1" (see Figure 3.2.3-2) (Illingworth *et al*, 2008; 2010). According to the "Ensemble Human" assembly from 2007 there are no genes, pseudogenes or other transcripts annotated within the 132.62 kb selected locus on human chromosome 18 (Flicek *et al*, 2012). Two homology arms, one 250 and the other 550 bp corresponding to regions in the middle of Gene desert 1 were amplified and cloned into a plasmid containing a neomycin selection cassette under a dual prokaryotic and eukaryotic promoter flanked by frt sites (Figure 3.2.3-1A). Subsequently the PuroGFP CGI-like sequence was cloned into that plasmid. The plasmid containing the homology arms, the CGI and a neomycin selection cassette was used to engineer a bacterial artificial chromosome (BAC) containing the identified Gene desert 1 by recombineering (Figure 3.2.3-1B).

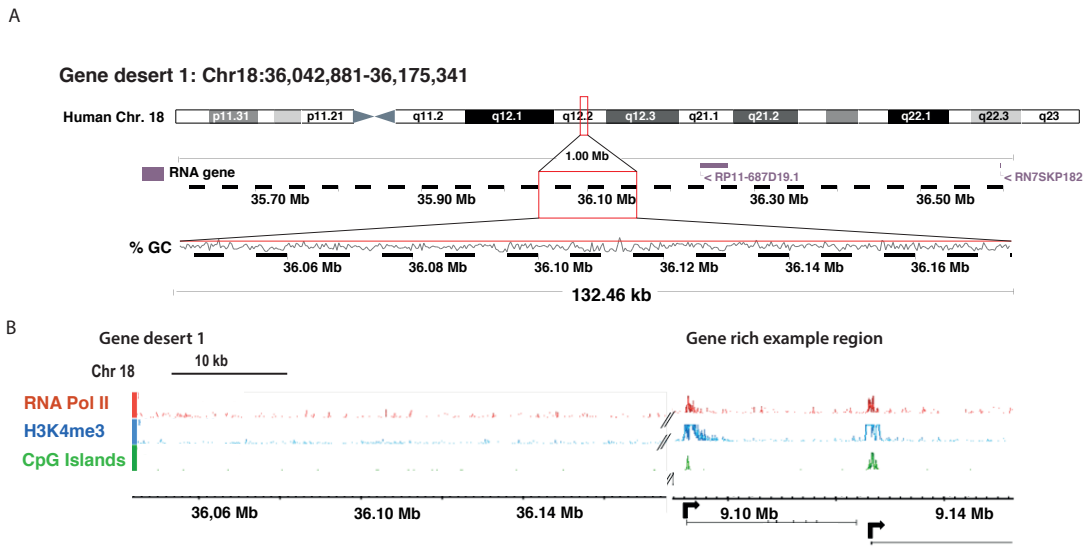


Figure 3.2.3-2 Genomic location of human Gene desert 1

A: Location of human Gene desert 1 on mouse chromosome 18. B: ChIP-Seq profiles for RNA Polymerase II and H3K4me3 and CAP seq profiles at Gene desert 1 and a gene rich example region for comparison.

3.2.3.2. Recombineering of CGI like sequences

Recombineering is a highly efficient and precise method for genetically engineering DNA *in vivo*, usually in *E. coli*, that relies on homologous recombination mediated by bacteriophage based systems such as the Red/Ed system. This method does not depend on restriction enzymes and is therefore more versatile than traditional cloning as the sequence of the homology region can be chosen freely. Moreover the size limit for fragments that can be recombineered is much less stringent than for other cloning methods and fragments up to 80 kb can be cloned into low-copy plasmids (Zhang *et al*, 1998). Initially the pRed/ET plasmid containing the phage protein pairs RecE/RecT or Redα/Redβ that are under the control of an arabinose inducible temperature sensitive promoter is electroporated into the BAC-containing bacteria. Double-stranded break repair (DSBR) is initiated by the recombinase protein pairs, where RecE or Red α, a 5'-> 3' exonucleases, digests one strand of the DNA from the DSB. The DNA binding proteins RecT/Redβ binds and coats the single strand and the protein-nucleic acid filament aligns with homologous DNA (Figure 3.2.3-3A). As a second step the linearized plasmid containing homology arms, a selection cassette and the PuroGFP is electroporated into the BAC containing the Red plasmid. Upon induction with arabinose recombination occurs and selection for successful recombination is achieved by plating bacteria on plates containing kanamycin (Figure 3.2.3-3B).

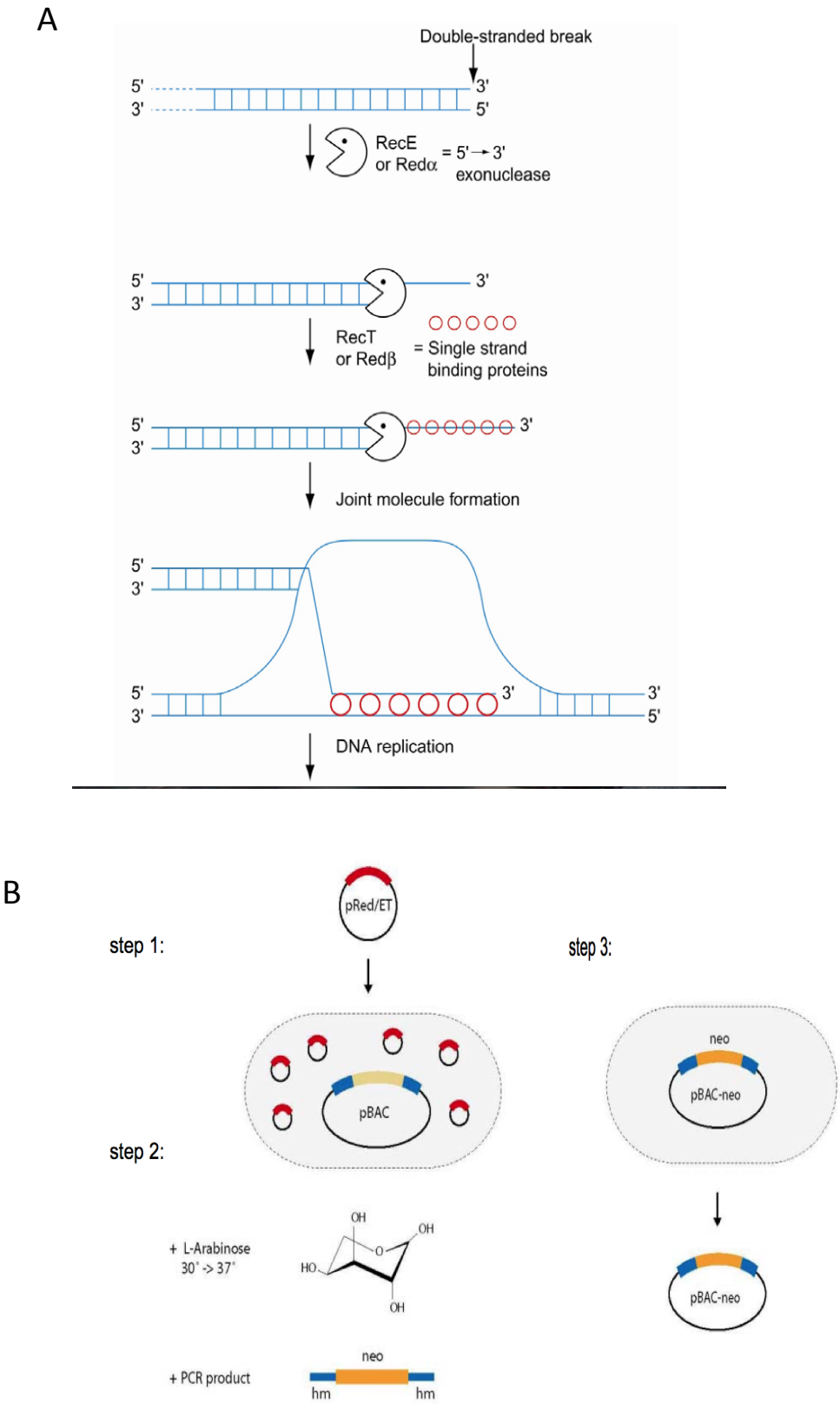


Figure 3.2.3-3 Schematic representation of principle of recombineering
A: Principle of recombineering. B: Stepwise description of recombineering. For description see main text. Adapted from (Stewart).

Clones were initially screened by PCR with primers spanning the 5' and 3' border and an internal primer pair for proper homologous recombination. Restriction enzyme digests with different enzymes (only *BamHI* shown) were performed to confirm the integrity of the whole 140kb BAC. As the size of the BAC is much bigger than conventional plasmids, achieving good band separation on an agarose gel proved to be difficult. However, despite the fact that some of the bands are visible only as one band in comparison with the *in silico* digest it is clear that the PuroGFP has been integrated (Figure 3.2.3-4A & B). Finally the integration was confirmed by sequencing along key elements of the BAC (5' and 3' border, PuroGFP and frt flanked neomycin cassette).

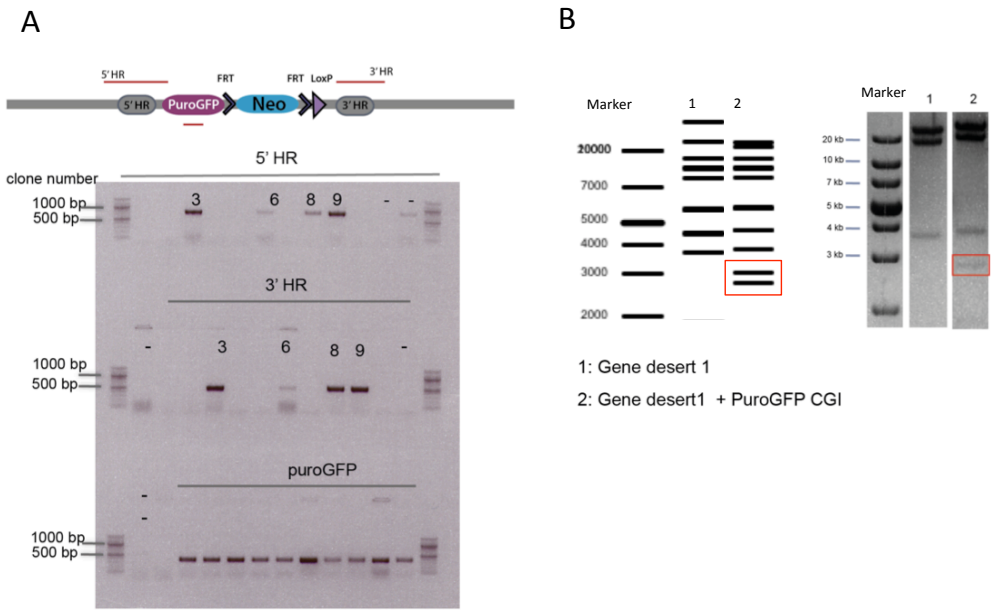


Figure 3.2.3-4 Introduction of CGI-like sequence into Gene desert 1 via recombineering
A: PCR screen to identify successful recombination. Red lines indicate amplicons that span 5'HR, 3'HR and an internal region. Numbers indicate clones that showed successful recombineering -: Water was used as a negative control. B: Control digest with *BamHI* to confirm recombination. *In silico* digest left panel, control digest right panel. Red box indicates diagnostic band.

3.2.3.3. Transfection of ES cells and excision of selection cassette

Mouse ES cells were transfected with the linearized BAC containing PuroGFP in Gene desert 1 leading to random integration. The ~ 70kb gene desert flanking puroGFP on either side provides insulator function protecting against positional effects. Successfully transfected cells were selected by addition of G418 and neomycin resistant colonies were picked and screened for the integration of the full length BAC using several PCR primers flanking the

modified BAC (see Figure 3.2.3-5). Since the integration of the BAC construct into the mouse genome happens randomly it is not possible to determine the proper integration by Southern blot, as is a usual approach for generating transgenic cell lines.

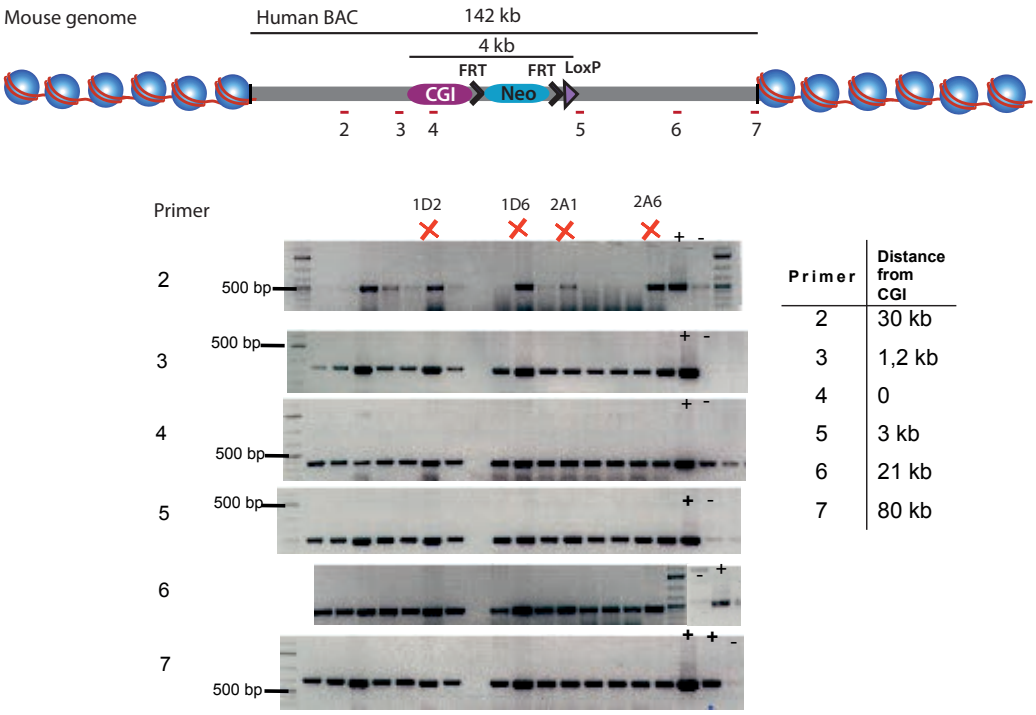


Figure 3.2.3-5 Screening of mouse ES cells for the integration of human gene desert BAC containing PuroGFP
Scheme on top indicates human gene desert BAC with PuroGFP CGI (grey bar) randomly integrated in genome (depicted by nucleosomes) of mouse ES cells. Red lines indicate primer amplicons. +: As a positive control BAC DNA from Gene desert 2 with the integrated CGI was used. -: As a negative control wt genomic DNA without transfected BAC was used. Table on the right denotes distance of primer from PuroGFP CGI. Red crosses indicate the clones that were chosen for selection cassette excision.

To eliminate the possibility of integration site effects, 3 different clones were expanded and electroporated with Flp in order to excise the selection cassette. This step is necessary in order to be able to analyse the chromatin at the CGI without an adjacent active promoter that could influence the result of the experiment. Resulting colonies were picked and initially screened by PCR using a primer pair that spanned the selection cassette. In the event of successful excision the size of the expected band drops from 2 kb to 0.5 kb. This screening method was chosen because a band is expected in either case, allowing assessing proper PCR conditions (Figure 3.2.3-6A&B). Clones that were positive for excision of the selection cassette were expanded further to perform Southern blot analysis. In contrast to PCR that relies on the amplification of DNA and is therefore prone to be biased, southern blotting does not require amplification. Here, the absence of the selection cassette can be shown by

the increase of the probed fragment from 1.5 to 12 kb (Figure 3.2.3-6 C&D). Figure 3.2.3-6 D shows clones that were positive in the PCR screen. However, only one of these was positive by Southern blot as well, proving the necessity of the Southern blot screen. As we aimed to analyse several cell lines clones with a different parental BAC integration were screened. Although several clones seemed positive by PCR, none had lost the selection cassette. Despite screening of more than 400 colonies from 3 different cell lines no additional clones with successful excision of the cassette were identified. When Southern blots were performed it became clear that all positive looking clones from the PCR screen (data not shown) were in fact mixed colonies (Figure 3.2.3-6 E). In summary, only one clone with the PuroGFP CGI-like sequence in gene desert 1 without the selection cassette was obtained for ChIP analysis.

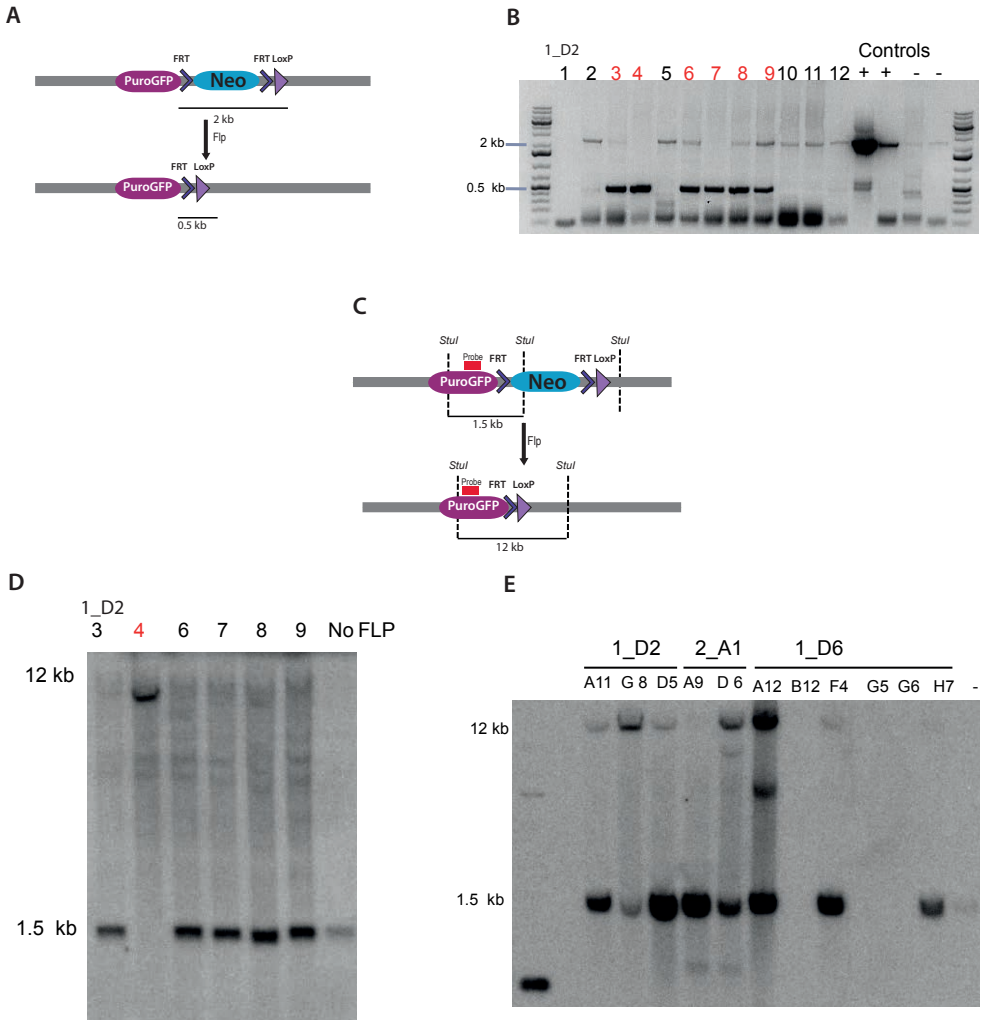


Figure 3.2.3-6 Excision of selection cassette

A: Scheme of PCR screen to identify clones without the selection cassette, indicated by a decrease of amplicon from 2kb to 0.5 kb (black line). B: PCR screen for the excision of the selection cassette. For Clones that were positive for the first crude screen (data not shown) the PCR was repeated (panel B). Clone numbers in red were taken further for Southern screen. C: Scheme of Southern screen. Red box denotes probe. Successful excision shown through increase of band from 1.5 kb to 12 kb. D: Clones from PCR screen. Red clone number indicates successful excision. E: Southern screen of more clones from different parental integrations (PCR screen not shown). Only mixed colonies were identified, no further clones of different parental integrations were obtained.

3.2.4. Both H3K4me3 and H3K27m3 marks are present at puroGFP in gene desert

In order to answer the question weather the PuroGFP CGI-like sequence used previously within genes is sufficient to create a bivalent domain outside transcriptional units the chromatin state of the one cell line was analysed by ChIP. Using an antibody against H3K4me3 we found that as in the work by Thomson and co-workers (Thomson *et al*, 2010) the promoterless PuroGFP sequence created a novel peak of H3K4me3. As can be seen in Figure 3.2.4-1 a defined peak was observed just over the CGI-like sequence that was not present in the adjacent gene desert regions. This suggests that the underlying CpG and G+C rich DNA influences local chromatin establishment. When looking at the level of H3K4me3 over the CGI it is notable that the height of the peak corresponds to the height of the bivalent gene *HoxC8*. As expected the H3K4me3 levels at an active gene like *ActinB* are higher. An inconspicuous region on mouse chromosome 15 was used as a negative control. Published genome wide ChIP data show that there is no histone H3 methylation at lysine 4 present at this region (Clouaire *et al*, 2012).

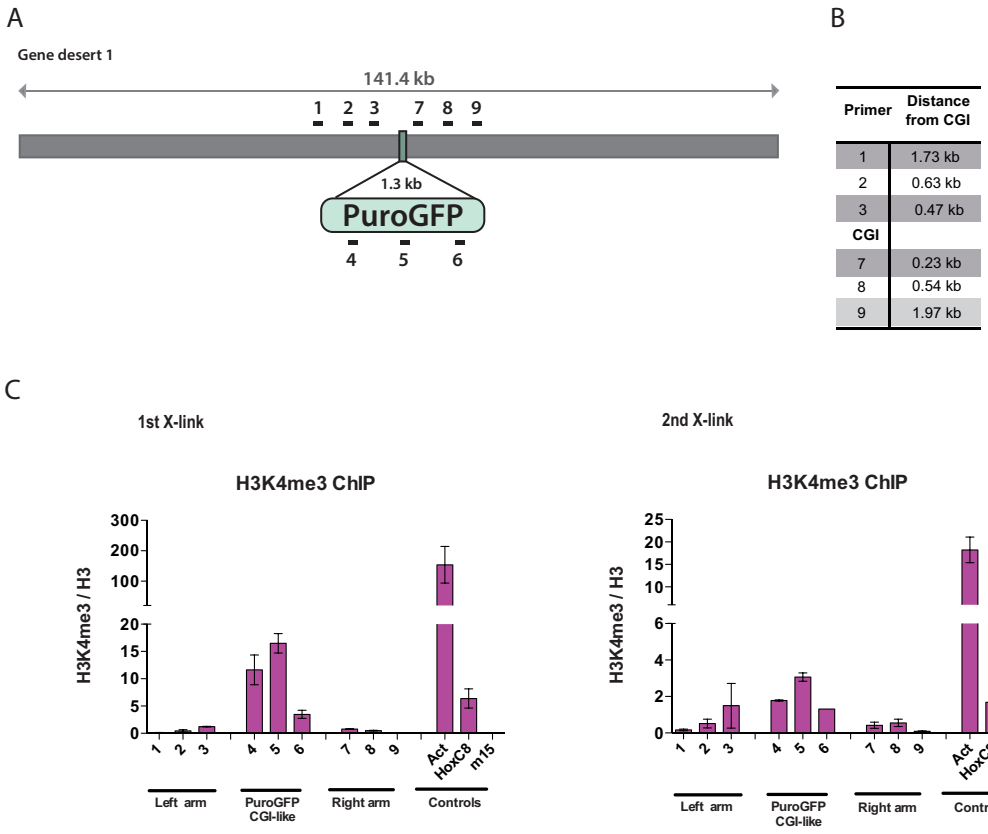


Figure 3.2.4-1 A novel peak of H3K4me3 is established over PuroGFP in Gene desert 1
A: Schematic showing the position of primer pairs used for ChIP. (length of amplicons not drawn to scale) B: Distance of primers from integrated CGI in kb. C: H3K4me3 ChIP, 2 independent X-links. Y-axis: % of Input of H3K4me3 over % of Input of pan-H3 antibody. Controls: TSS of active gene Actin, of bivalent gene HoxC8; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates.

Furthermore ChIP with antibodies against H3K27me3 was performed to investigate if this repressive chromatin mark is present at the inserted PuroGFP CGI-like sequence. Figure 3.2.4-2 B shows that H3K27me3 is established over the CGI-like sequence. In contrast to the H3K4me3 ChIP, this ChIP with an antibody against H3K27me3 shows a signal not only directly over the CGI but spreading into adjacent gene desert regions. This spreading of H3K27me3 has been observed by others (Schwartz & Pirrotta, 2007) and is thought to be caused by the ability of PRC2 to bind to the same mark it deposits, providing a model for the propagation of H3K27me3 during replication (Hansen *et al*, 2008; Margueron *et al*, 2009). It has to be noted that the primer pairs investigating the chromatin state at the gene desert region are quite close to the integration site of the CGI-like sequence. Therefore, it cannot be excluded that the gene desert region itself could have a role in establishing the H3K27me3 mark. For future experiments it will be interesting to analyse the presence of H3K27me3

further away from the integration site towards the end of the gene desert (see 4.2.1). Again the level of H3K27me3 at the CGI-like sequence is similar to that of the bivalent gene *HoxC8*. In order to strengthen the evidence that H3K27me3 is indeed present at the CGI, ChIP with antibodies against Suz12, a subunit of the PRC2 complex was performed. Panel C of Figure 3.2.4-2 indicates that Suz12 is recruited to the PuroGFP CGI-like sequence but is absent from adjacent gene desert regions. As before the levels are comparable to that of *HoxC8*. These results indicate that the PuroGFP sequence that contains a high frequency of CpGs and a high G+C content is sufficient to create a bivalent domain independent of the neighbouring sequence.

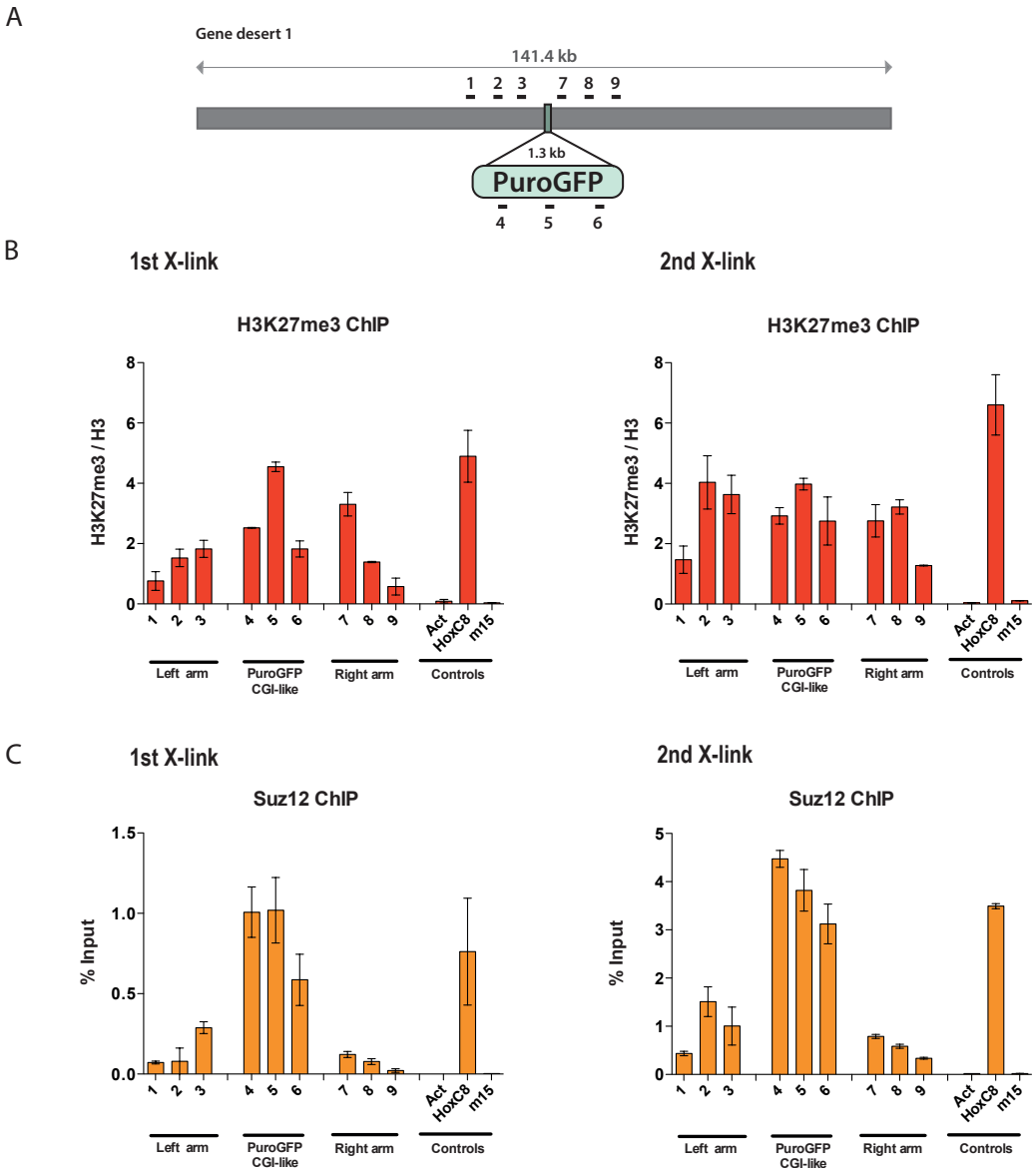


Figure 3.2.4-2 Polycomb is recruited to PuroGFP in Gene desert 1

A: Schematic showing the position of primer pairs used for ChIP. (length of amplicons not drawn to scale) B: H3K27me3 ChIP in 2 independent X-links. Y-axis: % of Input of H3K4me3 over % of Input of pan-H3 antibody. C: Sux12 ChIP, Y-axis % Input. Controls: TSS of active gene Actin, of bivalent gene HoxC8; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates.

3.3. Summary and Discussion

CpG islands are short stretches of unmethylated DNA with a high CpG density and an overall high G+C content that punctuate the globally heavily methylated and CpG poor bulk genome. Moreover they are associated with the TSS of many genes, which adds to the notion that these islands are functionally important sequences. However, many questions remain regarding their active influence on processes like chromatin structure establishment. In this chapter earlier findings that indicated promoterless CGI-like sequences can recruit a histone methyltransferase complex and establish a novel domain of H3K4me3 (Thomson *et al*, 2010) were confirmed and extended. The establishment of the active chromatin mark H3K4me3 over the inserted CGI-like sequence was confirmed for both cases: the Nanog-puroGFP insertion and the Mecp2-eGFP construct

3.3.1. Polycomb is recruited to a promoterless CGI-like sequence at the *Nanog*-PuroGFP and *Mecp2*-eGFP loci

Polycomb group repressor proteins play a crucial role in regulating gene expression patterns through differentiation by conferring a repressive chromatin environment at target genes. The exact recruiting mechanism of polycomb proteins has not been completely clarified. In *Drosophila* the recruitment of PcG complexes is mediated by specific DNA elements called Polycomb response elements (PREs). Sequence specific DNA binding proteins, such as PHO, recognize PREs and recruit polycomb proteins (Ringrose & Paro, 2007). Some reports suggested the existence of a mammalian PRE and implied a role for YY1, the mammalian orthologue of the *Drosophila* protein PHO, in Polycomb recruitment (Lo *et al*, 2012; Sing *et al*, 2009). Yet genome-wide analysis in mammals did not show a clear overlap between YY1 and PcG target genes (Xi *et al*, 2007). When looking at genome wide studies of Polycomb occupancy, it becomes apparent that there is a strong correlation of Polycomb binding to promoters with CGIs (Ku *et al*, 2008). Recent studies suggest that the accessible chromatin state usually found at CGIs is owed at least partially to the zinc finger-CxxC domain-containing proteins that specifically recognize and bind to unmethylated CpGs and recruit chromatin modifying activities (Thomson *et al*, 2010; Blackledge *et al*, 2010). Given the fact

that most polycomb proteins are recruited to CGIs it is interesting to speculate if a similar mechanism plays a role for polycomb targeting. To date a clear preference for unmethylated CpGs has not been demonstrated for the PRC2 complex but Jarid2, a member of this complex, has been shown to preferentially bind to GC rich DNA sequences (Peng *et al*, 2009; Li *et al*, 2010). It has been suggested that CG rich DNA is sufficient to recruit polycomb proteins (Mendenhall *et al*, 2010; Lynch *et al*, 2011).

Indeed it is shown here that the H3K27me3 histone mark that is established by a member of the PRC2 complex is present at the promoterless CGI-like sequence GFP and puroGFP that have been integrated into the 3'UTR of Mecp2 and Nanog, respectively. This suggests that a sequence with the characteristics of a CGI but stemming from an unrelated organism is capable of recruiting the polycomb complex to locations previously not marked with this modification.

3.3.2. Introduction of CGI-like sequences into a gene desert region

In this chapter the above-discussed finding within genes were extended to a gene desert region. To this end only the PuroGFP CGI-like sequence was taken further and a cell line was created by random integration that contained the PuroGFP sequence in a gene desert region. As each construct will integrate into a different genomic locus making comparisons potentially challenging, measures where taken to prevent negative effects of the random integration approach. Long genomic stretches of human gene desert BACs either side of the integrated CGI-like sequence provide insulator function to protect against positional effects. The idea to create several cell lines of the gene desert 1 with puroGFP all integrated into different genomic loci failed due to inefficient excision of the selection cassette. The construct used in this study to confer antibiotic resistance was an frt flanked neomycin gene under a PGK promoter. The frt sites are recognized by the recombinase Flp, which leads to the excision of the selection cassette. It has been shown that the efficiency of Flp mediated recombination is less than that achieved with the traditional Cre recombinase that recognizes LoxP sites (Ringrose *et al*, 1998). For this study it was not possible to use the Cre/LoxP system because the bacterial backbone of the human gene desert BAC contains one remaining LoxP site. Undesirable recombination events could occur between this residual LoxP site and the LoxP sites flanking the neomycin gene that could potentially lead to excision of the CGI. It was therefore necessary to use the less efficient Frt/Flp system. However, for different constructs the protocol for excision of the selection cassette was improved to increase the efficiency of excision (see chapter 4.2.1).

3.3.3. Polycomb levels at CGIs integrated in gene desert are higher than in active genes

In this chapter the influence of the promoterless CGI-like sequence PuroGFP on chromatin establishment was analysed in the 3'UTR of the Nanog gene and in a human gene desert. The levels of H3K27me3 over the insertion within the Nanog gene are markedly lower than that of the bivalent control gene HoxC8 but still clearly above background (Figure 3.2.1-1). In contrast when looking at the same insertion in a gene desert region, where there are no influences of neighbouring genes expected, the H3K27me3 intensity is higher in both replicates reaching levels similar to that of HoxC8 (see Figure 3.2.4-2). The higher levels of H3K27me3 correspond to higher levels of Suz12, a component of the PRC2 complex. Whereas Suz12 levels at the insertion within the Nanog locus was around 5 times lower than that of the bivalent control gene. This result is not surprising giving the fact that the Nanog gene is highly expressed in mouse ES cells in order to maintain pluripotency. It is therefore conceivable that the recruitment of Polycomb to a CG rich sequence within an active gene, albeit at the 3'UTR, is restricted and counteracted by activating processes. This is consistent with recently emerging ideas about polycomb recruitment that favour the view that PcG proteins are constantly sampling theoretically favourable binding sites, mostly CG rich sequences, for permissiveness to bind. If favourable conditions are found then binding of PcG is enhanced through positive feedback mechanism. On the other hand processes such as transcription, can lead to the domination of activating signals by recruiting activating histone-modifying enzymes. Mendenhall and co-workers showed that a constitutively active CpG island is able to recruit PRC2 after excision of a cluster of activating motifs (Mendenhall *et al*, 2010). This view was confirmed by a study that showed competition between PcG recruitment and transcriptional activation by inserting or excising a small promoter element and showing that the presence of the promoter is incompatible with PcG binding (Lynch *et al*, 2011). Low gene expression might be already enough to inhibit full PcG establishment as the GFP CGI-like sequence integrated at the 3'UTR of Mecp2, a gene that is expressed at very low levels in ES cells, shows similar levels of H3K27me3 and Suz12 to the Nanog-PuroGFP integration. However, this remains speculative and more examples are needed to be able to draw conclusions about quantitative levels of H3K27me3 in different locations.

3.3.4. H3K4me3 is created at basal levels over CGI-like sequence but transcription is needed for full establishment

In contrast to H3K27me3, H3K4me3 levels of the puroGFP construct within a gene desert region are similar to the levels observed when this sequence was integrated into the 3' UTR of the *Mecp2* and *Nanog* genes. In both scenarios H3K4me3 signals were found to be around the height of the bivalent control gene HoxC8, but much less compared to the levels observed at the TSS of active genes. This indicates that although there is a basal level of H3K4me3 recruited to promoterless CGIs, presumably via a CxxC domain-mediated targeting mechanism, the levels of H3K4me3 are markedly increased by productive transcription. Several studies have high lightened the importance of RNA polymerase II dependent recruiting mechanism of H3K4 methyltransferases (Guenther *et al*, 2007; Lee & Skalnik, 2008; Ng *et al*, 2003). For example, Wdr82 recruits the Setd1A histone methyltransferase complex to TSSs of transcribed genes via the interaction with the Ser5 phosphorylated form of RNA PolII (Lee *et al*, 2007a).

In summary it can be concluded that the promoterless CGI-like sequence puroGFP can influence local chromatin establishment by recruiting histone-modifying enzymes that form a bivalent chromatin domain over the integration site, indicated by the presence of both H3K4me3 and H3K27me3.

4. A artificial CGI-like sequence is sufficient to establish bivalent chromatin in a gene desert region

4.1. Introduction

4.1.1. Establishment of bivalent domains

Bivalent domains that are marked by the coexistence of the permissive chromatin mark H3K4me3 and the repressive mark H3K27me3 are thought to play an important role by keeping developmental genes silenced but poised for activation upon differentiation (Azuara *et al*, 2006; Bernstein *et al*, 2006). However, many questions remain regarding their establishment and biological significance. Especially the notion that key developmental genes are specifically targeted by PcG and Set containing proteins to establish a bivalent domain might be misleading as it has been shown in earlier studies that CG rich sequences are sufficient to recruit the PRC2 complex (Mendenhall *et al*, 2010; Lynch *et al*, 2011) and an H3K4 methyltransferase (Thomson *et al*, 2010). These results have been recapitulated in chapter 3, where it was shown that the promoterless PuroGFP CGI-like sequence formed a bivalent domain in a gene desert. Therefore it is conceivable that a bivalent domain is the default state of a CpG rich sequence with a high G+C content in the absence of any other activating or repressing cues.

4.2. Results

4.2.1. Both H3K4me3 and H3K27m3 marks are present at artificial CGI while RNA Polymerase II is not detected

To demonstrate that the observed results are not due to some specific characteristic of the PuroGFP sequence, which was created from the coding region of two exogenous genes, but constitute a more general mechanism of bivalent domain formation at CpG and G+C rich sequences a synthetic CGI-like sequence was analysed. This sequence was generated randomly with the constraint to have similar parameters to endogenous CGIs in terms of length, number of CpGs and G+C content. This artificial CGI like sequence (ArtCGI) is 1055 bp long, has a G+C content of 69.5 % and 12 CpGs per 100 bp. Moreover, care was

taken to avoid the core consensus motive of the ubiquitous transcription factor Sp1 (CCGCCC). This was deemed important because earlier studies have found that Sp1 sites are required to keep the hamster and mouse *aprt* gene, respectively, promoter methylation free (Brandeis *et al*, 1994; Macleod *et al*, 1994). Macleod et al showed that mutating Sp1 sites leads to methylation of the *aprt* CGI. As I want to study the influence of the base composition of CGIs on local chromatin establishment, the presence of a transcription factor binding site that could influence the methylation status of the CGI and therefore presumably chromatin formation was considered detrimental. However, since CGIs are GC rich and are located at the promoter regions of genes they are naturally rich in transcription factor binding sites (Qian *et al*, 2006). Therefore other transcription factor binding sites are likely to be present in the synthesized artificial CGI.

As described in the previous chapter the artificial CGI was cloned into the plasmid harbouring the homology arms for recombineering into the human Gene desert 1 BAC with a neomycin resistance cassette (Figure 3.2.3-1). However, the recombineering step into the 132kb big BAC containing Gene desert 1 proved to be too challenging. Despite many attempts no correct recombination events were identified, presumably due to additionally recombination events within repetitive region in the gene desert. Care was taken to avoid repetitive regions as it was anticipated that the presence of those elements could potentially interfere with homologous recombination. Nevertheless some degree of repetitiveness could not be avoided. In order to circumvent this problem a different human gene desert region was selected that is smaller than the one previously used. We chose to obtain a BAC (BAC-L19) containing a 47 kb region on human chromosome 1. This region has been used for a similar approach (Mendenhall *et al*, 2010) therefore we expected it to be favourable to engineering. Figure 4.2.1-1 shows the genomic location and ChIP Seq profiles of H3K4me3 and RNA Pol II as well as CAP-Seq of gene desert 2.

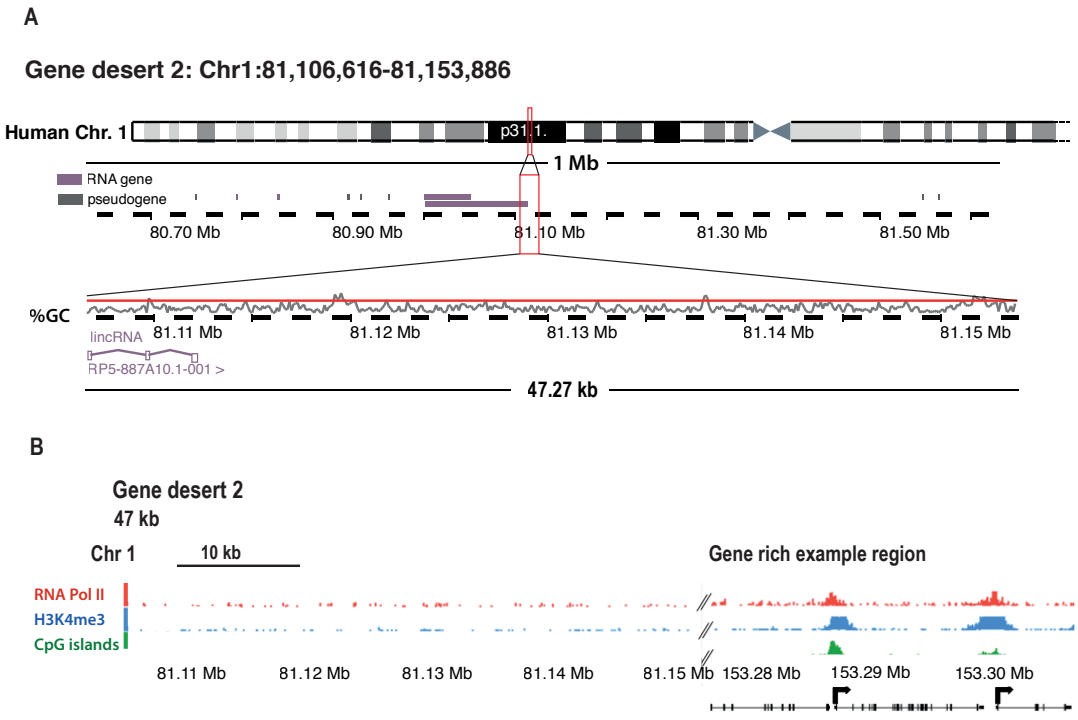


Figure 4.2.1-1 Genomic location of gene desert 2
A: Location of human gene desert 2 on mouse chromosome 1. B: ChIP-seq profiles for RNA Polymerase II and H3K4me3 and CAP-seq profiles at gene desert 2 and a gene rich example region for comparison.

As hoped, recombineering of the artificial CGI into gene desert 2 proved to be more efficient. Again mouse ES cells were transfected with the linearized BAC containing the selection cassette and the artificial CGI. Clones resistant to G418 were picked and screened for the integration of the construct. The number of BAC molecules integrated into the mouse genome was one concern that was addressed by determining transgene copy number by quantitative PCR. Figure 4.2.1-2 shows the measured copy number for several clones. Three clones that showed a low copy number integration for all 3 primer pairs (1, 1-2 and 4 respectively) were expanded further in order to excise the selection cassette.

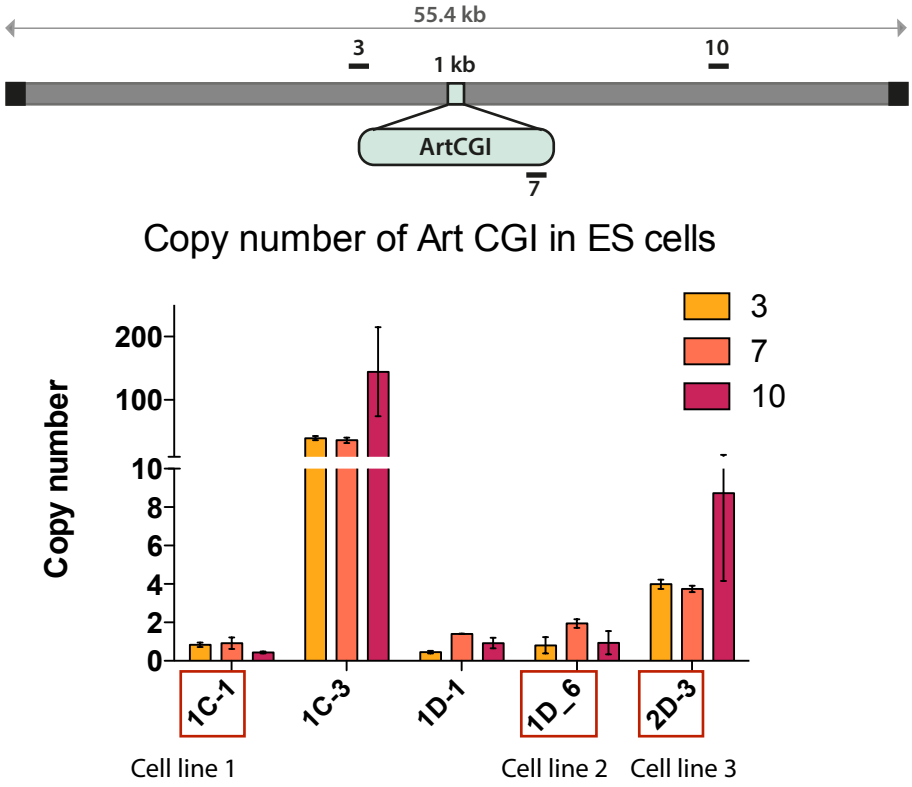


Figure 4.2.1-2 Copy number of artificial CGI in mouse ES cells in gene desert 2
Mouse ES cell clones with integrated artificial CGI were analysed by Q-PCR for copy number integration. 3, 7 and 10 respectively refers to the primer pair used in the PCR. Values normalized to that of Sox2 = 2 copies per cell. Red boxes indicate chosen cell lines. Error bars indicate standard deviation of PCR triplicates. Scheme above indicates construct inserted in gene desert 2 and position of primers (not drawn to scale).

Engineering the PuroGFP cell line showed that the limiting step in the procedure is identifying clones that have undergone successful excision of the selection cassette. Therefore I tried to optimize the Flp transfection protocol in order to increase the number of positive clones. This was achieved by adding an additional selection step that ensured only cells that did uptake the Flp plasmid were able to survive. Applying a transient selection with puromycin increased the number of positives by PCR screen from around 5/100 to more than 30/100 (see Materials and Methods, section 2.2.4). Figure 4.2.1-3 panel A shows the PCR screen after the improved protocol. Positive clones were taken forward and Southern blot analysis was performed. Figure 4.2.1-3 B shows positive clones for all three cell lines.

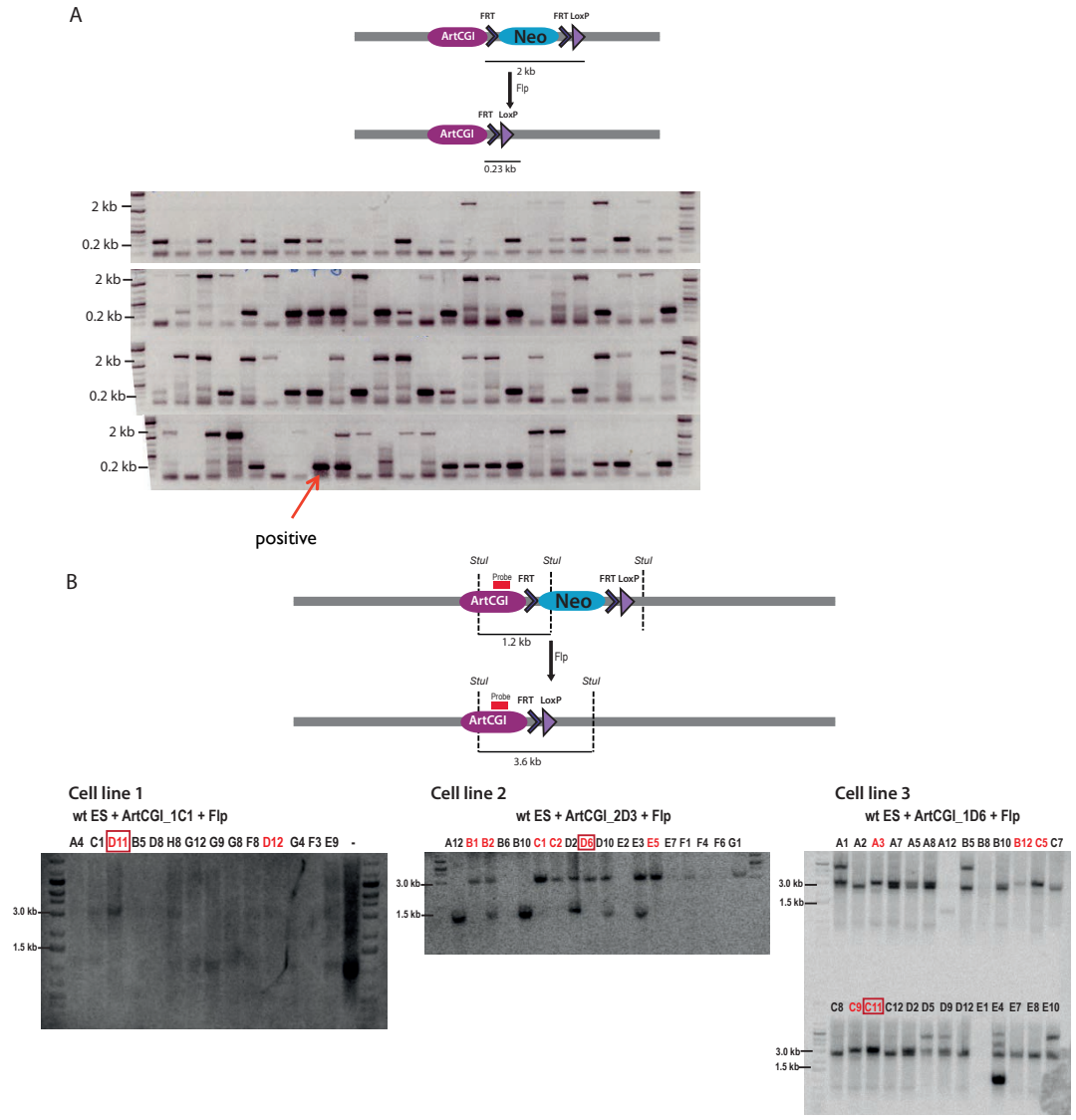
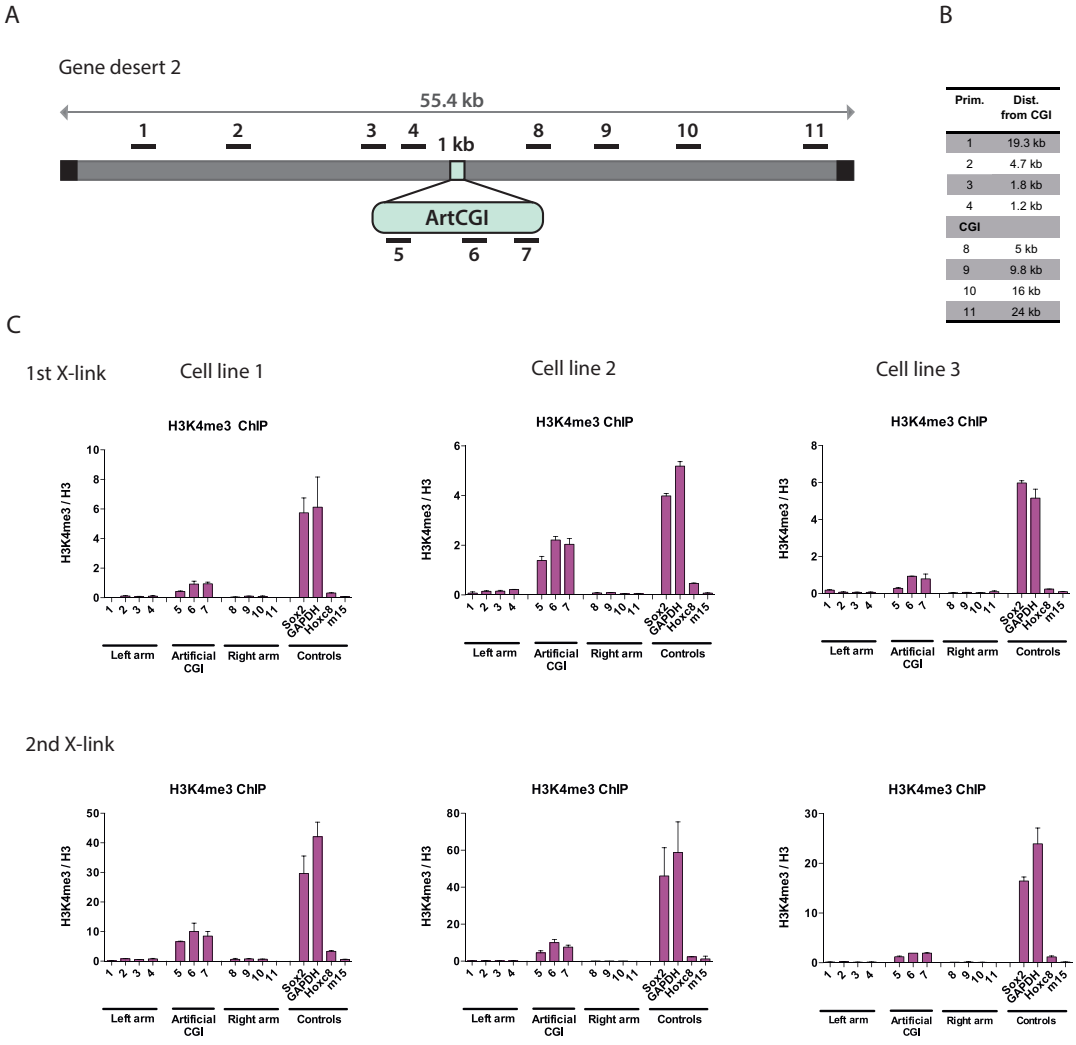


Figure 4.2.1-3 Identification of cell lines with excised selection cassette
A: PCR screen: Black line at bottom of scheme denotes amplified region. Red arrow indicates an example positive clone. B: Southern Blot screen: Red filled box indicates location of probe. Clones that depict a 3.6 kb band = excision of cassette are indicated in red. Red box denotes which clones were taken further for ChIP analysis

Three independent cell lines containing the artificial CGI in a gene desert region were analysed with antibodies against H3K4me3 in two independent experiments. The ChIPs in Figure 4.2.1-4 show the expected patterns at the control regions, high levels of H3K4me3 at the active genes *Sox2* and *GAPDH*, lower levels at the bivalent gene *HoxC8* and no enrichment over the negative control region m15. As with the PuroGFP the artificial CGI established a novel mark of H3K4me3 over the integrated sequence. Histone H3K4 methylation levels were found in all three cells lines to levels similar or slightly higher than at the bivalent gene *HoxC8*. In order to investigate the chromatin state at the gene desert,

primer pairs were designed spanning the whole gene desert length. The table in Figure 4.2.1-4 B indicates the distance of the specific primers from the integrated artificial CGI. No enrichment of H3K4me3 was found at the gene desert region demonstrating that it is specifically the randomized CpG and G+C rich sequence that is responsible for the establishment of the active chromatin mark H3K4me3.



Prim.	Dist. from CGI
1	19.3 kb
2	4.7 kb
3	1.8 kb
4	1.2 kb
CGI	
8	5 kb
9	9.8 kb
10	16 kb
11	24 kb

Figure 4.2.1-4 An artificial CGI-like sequence establishes a novel peak of H3K4me3
A: Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars indicate position of primers used for Q-PCR (length of amplicons not drawn to scale). B: Table shows the distance from each gene desert primer pair from the CGI in kb. C: H3K3me3 ChIP, 2 independent X-links for 3 cell lines. Y-axis: % of Input of H3K4me3 over % of Input of pan-H3 antibody. Controls: TSS of active genes Sox2 and GAPDH, of bivalent gene HoxC8; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates.

Again H3K27me3 ChIPs were performed to establish if the artificial CGI promotes the creation of a bivalent domain. Figure 4.2.1-5 shows that as with the PuroGFP CGI-like

sequence, H3K27me3 is found over the artificial CGI spreading into adjacent gene desert regions. This is despite the fact that for this experiment gene desert primers were designed up to 24 kb away from the integration site. However, a decline of H3K27me3 levels can be observed the further away from the integration site the primer pair is located. Additionally Suz12 is recruited to the artificial CGI but not to the adjacent gene desert region further strengthening the notion that a bivalent CGI has been created. Having shown that a bivalent domain is established over the artificial CGI in a gene desert we wanted to further characterize the chromatin state at the inserted sequence.

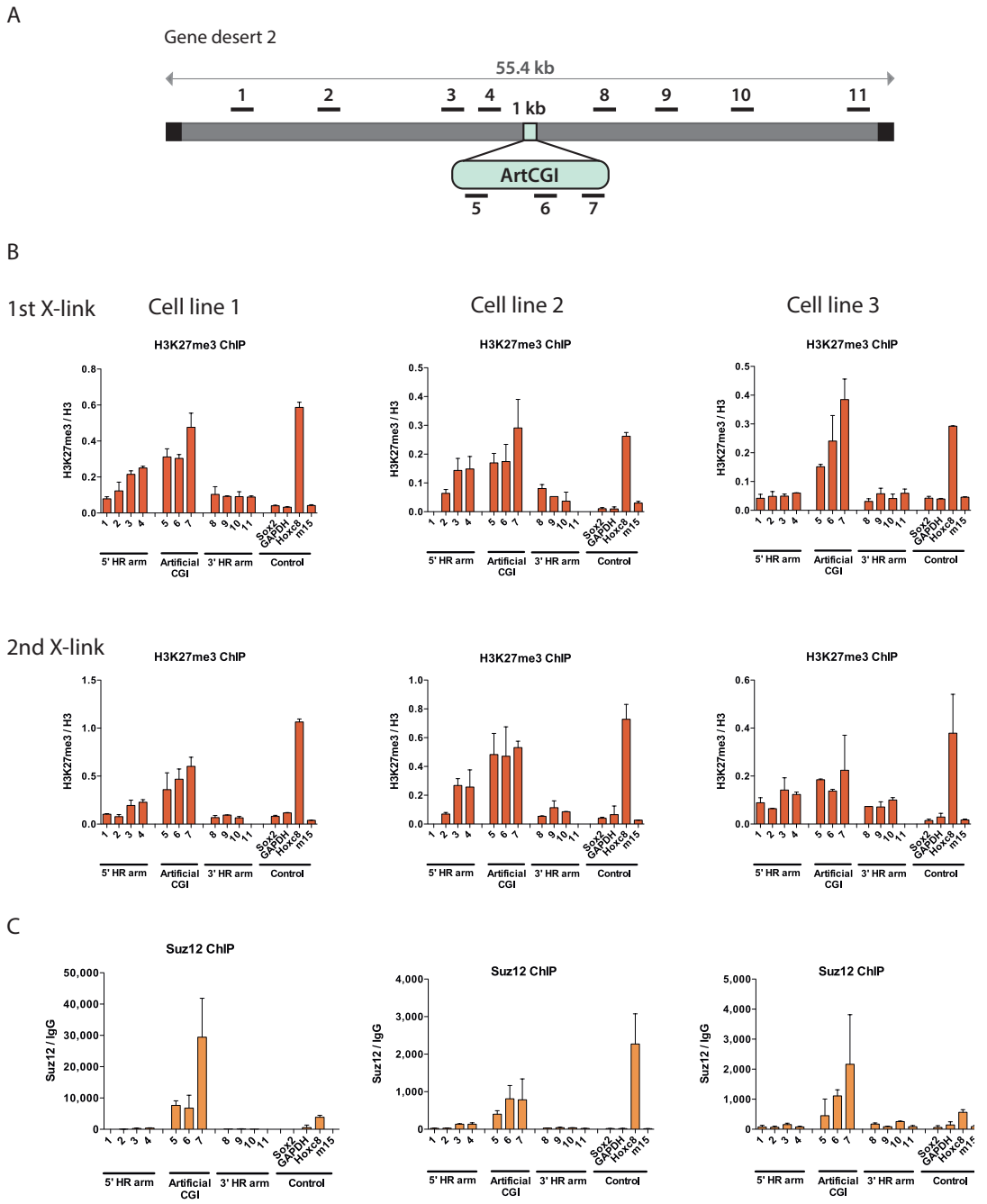


Figure 4.2.1-5 H3K27me3 and Suz12 ChIP of artificial CGI in gene desert 2
A: Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars indicate position of primers used for Q-PCR (length of amplicons not drawn to scale). B: H3K427me3 ChIP, 2 independent X-links for 3 cell lines. Y-axis: % of Input of H3K4me3 over % of Input of pan-H3 antibody and C: Suz12 ChIP, 1 X-link for 3 cell lines. Y-axis: % input. Controls: TSS of active genes Sox2 and GAPDH, of bivalent gene HoxC8; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates.

Originally, H3K4me1 was thought to be enriched only at enhancer elements (Heintzman *et al*, 2007). However, in recent years it has become apparent that H3K4me1 is not restricted to enhancers but is also present at broad regions of 5' ends of actively transcribed genes and even in some cases at noncoding sequences (Calo & Wysocka, 2013). Interestingly, there exists a subclass of enhancer, so called poised enhancers that are predominantly found in pluripotent ES cells and are marked by the coexistence of H3K4me1 and H3K27me3 together with members of the PRC2 complex (Rada-Iglesias *et al*, 2011). Therefore, we were interested to analyse the artificial CGI for presence of H3K4me1. Performing ChIP with an antibody against H3K4me1 we found an enrichment of this mark at the artificial CGI compared to the negative control region m15. The levels were similar to that at the transcription start site of *HoxC8* and slightly higher than at the TSS of *Sox2* (Figure 4.2.1-6). It will be interesting in the future to include a primer pair located to an enhancer region in the experiment in order to get information about the relative levels at the CGI in comparison to a region where H3K4me1 is expected to be highly enriched.

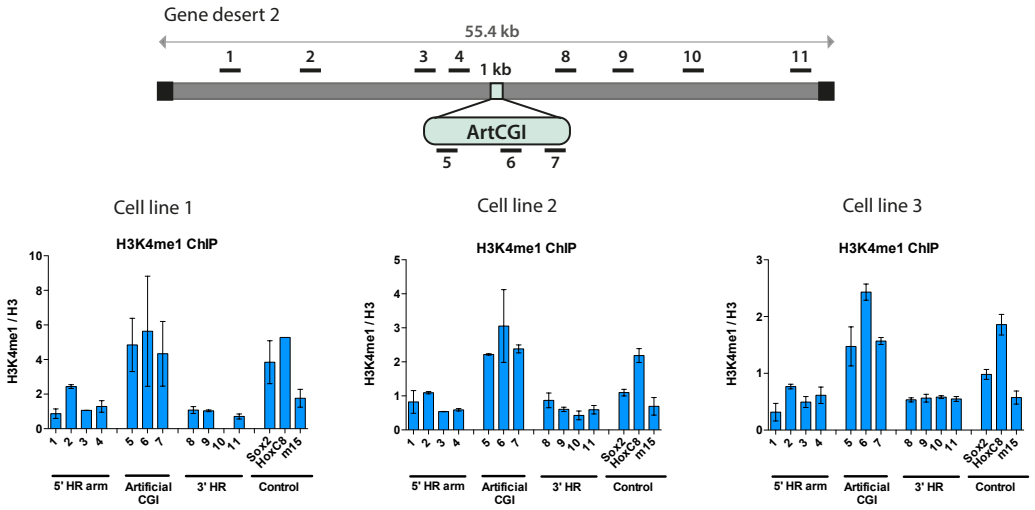


Figure 4.2.1-6 H3K4me1 ChIP of artificial CGI in gene desert 2
Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars indicate position of primers used for Q-PCR (length of amplicons not drawn to scale): H3K3me1 ChIP for 3 cell lines. Y-axis: % of Input of H3K4me3 over % of Input of pan-H3 antibody. Controls: TSS of active gene Sox2, of bivalent gene HoxC8; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates.

As it is has been shown that the transcriptional machinery is able to recruit H3K4me3 methyltransferases to chromatin, the establishment of H3K4me3 over the artificial CGI might be a by-product of transcription and not an intrinsic feature of the base composition of

CGIs (Ruthenburg *et al*, 2007). The majority of CGIs co-localize with H3K4me3, RNA Pol II and Cfp1 as was determined by ChIP-Seq in our laboratory (Thomson *et al*, 2010). There is, however, a small percentage of CGIs (7 %) that are characterized by lack of PolII despite the fact that there are prominent peaks of H3K4me3 and Cfp1 present (Thomson *et al*, 2010). This raised the possibility that in some cases RNA Pol II is not required and that the underlying CGI sequence is sufficient for the establishment of H3K4me3. This notion was supported by the fact that the promoter-less CGI like sequence created a novel peak of H3K4me3 but no RNA Pol II was detected. We were therefore interested to investigate if RNA Pol II is recruited to the artificial CGI. Figure 4.2.1-7 shows the ChIP results for all three cell lines for two independent replicate experiments. Panel A shows the ChIP results with an antibody that recognizes the N-terminus of the RNA polymerase and does therefore not distinguish between the phosphorylation states of the C-terminal domain (CTD). From the first experiment it became clear that no RNA Pol II was detected over the artificial CGI in any of the cell lines. As expected there were clear signals detected at the TSS of active genes such as Sox2 and GAPDH. In the 2nd experiment cell line 2 shows a signal over the insertion. However, when looking at the negative control region m15, this region shows a high signal as well, indicating that this ChIP shows higher background levels than normal and that the detected Pol II signal is likely to be noise rather than true enrichment. In order to strengthen the finding 2 additional antibodies were used that recognize the unphosphorylated form of the CTD and the serine 5 phosphorylated one respectively. Similarly no RNA Pol II signal, or only levels comparable to that of m15, was detected over the artificial CGI with these antibodies (see panel B and C of Figure 4.2.1-7).

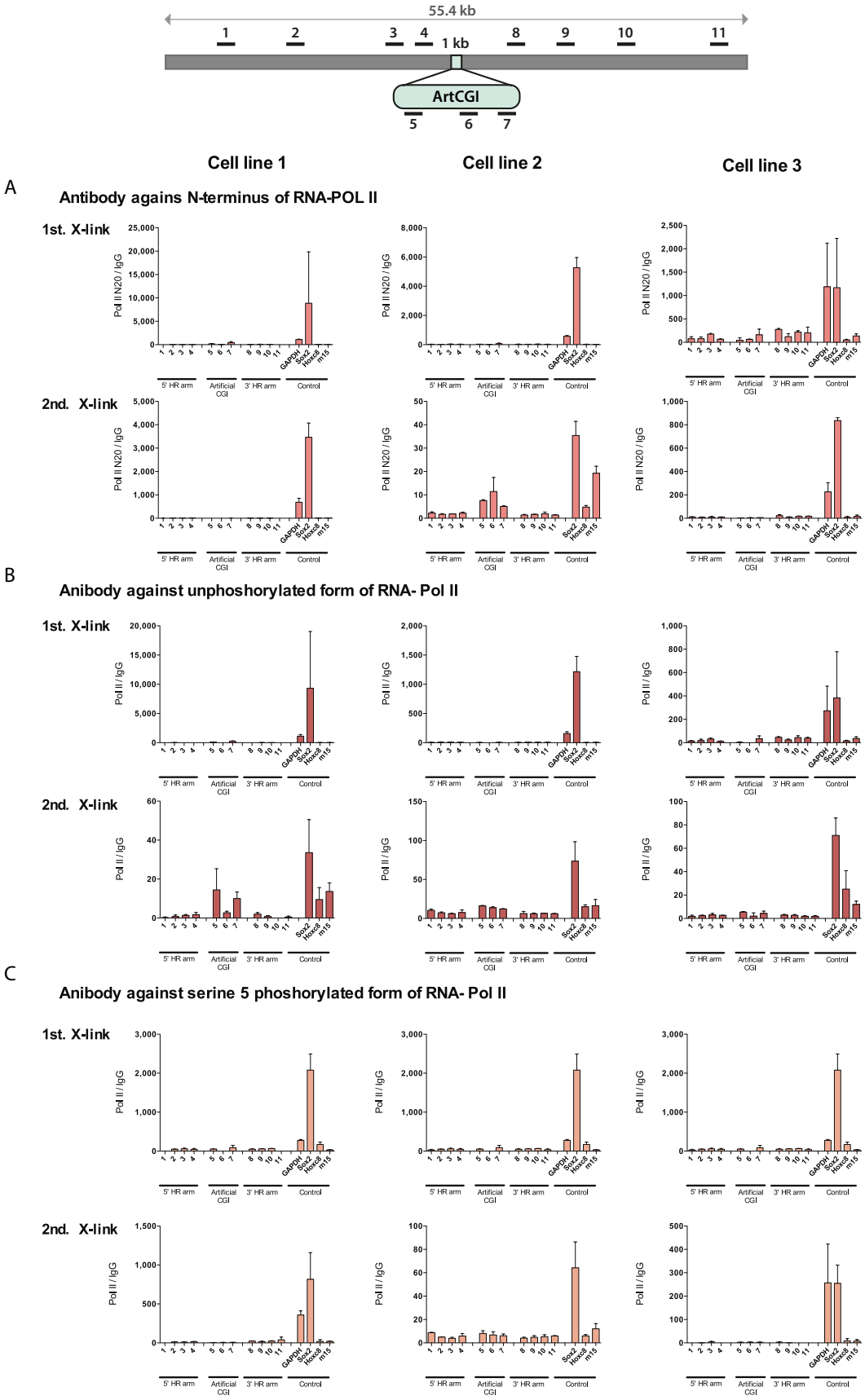


Figure 4.2.1-7 RNA polymerase II is not detected at artificial CGI in gene desert 2

Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars donate position of primers used for Q-PCR (length of amplicons not drawn to scale). A: PolII-N20 ChIP, B: PolII- unphosphorylated, C: PolII S5P for all three cell lines in 2 biological replicates. Y-axis: % of Input. Controls: TSS of active genes Sox2 and GAPDH, of bivalent gene HoxC8; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates.

Fitting with the absence of RNA Pol II, acetylated histones levels were low over the artificial CGI. High levels of acetylated histone 3 and 4 are usually associated with H3K4me3, gene promoters and high levels of gene expression. However, as with many histone modifications the exact role and way of establishment remain elusive. A recent study confirms the high correlation of H3K9/K14 acetylation in mouse ES cells with H3K4me3. Moreover, the levels of H3K9/K14ac directly correlate with the CpG content of the promoters attesting the importance of sequences underlying specific histone modifications (Karmodiya *et al*, 2012). They also provide evidence that those marks occur at bivalent promoters. Figure 4.2.1-8 shows ChIP experiments using an antibody against H3K9/K14 acetylation. In all cell lines H3K9/K14 acetylation levels seemed to be similar to that of the bivalent gene HoxC8. In the case of the artificial CGI in gene desert it seems that high CpG density is sufficient to recruit HATs to low levels, maybe through interaction with H3K4me3. However, the presence of the transcription machinery is likely to be necessary in order to establish full levels of acetylated histones. In summary we conclude that an artificial CGI like sequence influences local chromatin structure and establishes a bivalent domain. As expected for a construct that lacks a proper promoter sequence no Polymerase was detected and accordingly only low levels of acetylation were present.

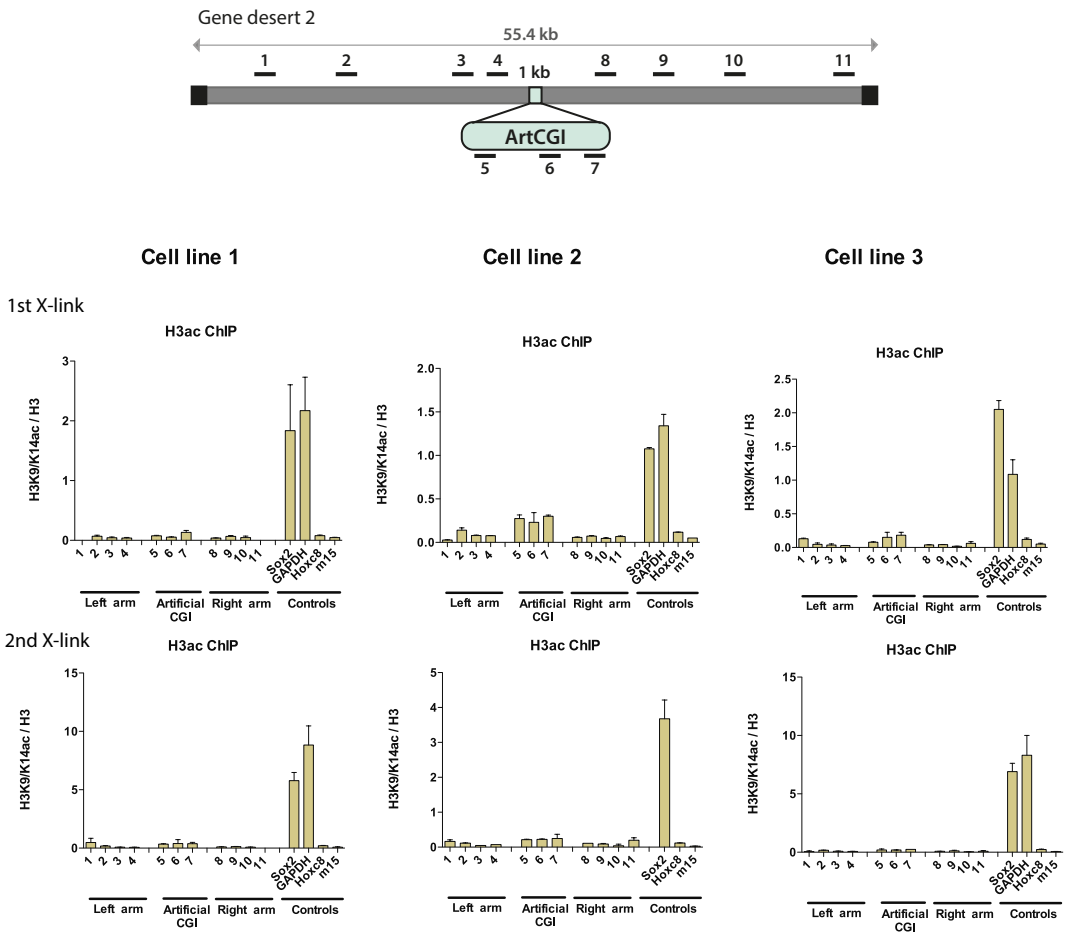


Figure 4.2.1-8 Low levels of H3K9/K14ac established over artificial CGI in gene desert
Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars indicate position of primers used for Q-PCR (length of amplicons not drawn to scale): H3K9/K14ac ChIP for 3 cell lines in 2 biological replicates. Y-axis: % of Input of H3K4me3 over % of Input of pan-H3 antibody. Controls: TSS of active gene Sox2, of bivalent gene HoxC8; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates.

4.2.2. An artificial CGI-like sequence is enough to be protected from DNA methylation in mouse ES cells

As it is unclear if CpG clusters are sufficient to stably perpetuate non-methylated CpG islands or if transcription is required additionally to maintain the non-methylated state, we analysed the methylation status of the introduced artificial CGI in gene desert 2. Figure 4.2.2-1 shows that all three lines display only low levels of methylation. This is on one side surprising as some previous studies found that CpG rich clusters become methylated in absence of transcription. For example point mutations that prevent transcription factor binding without significantly reducing CpG density destroy the immunity of a CpG island to

DNA methylation (Brandeis *et al*, 1994; Macleod *et al*, 1998). When CpG rich sequences from the *E.coli* genome were inserted into the mouse gene most sequences became heavily methylated (Lienert *et al*, 2011). Also more than half of cells carrying the promoter-less eGFP insertion at the *Mecp2* locus had acquired dense methylation in ES cells despite the presence of a CpG cluster (Thomson *et al*, 2010). On the other side there have been reports that histone methylation on lysine 4 interferes with Dnmt activity by inhibiting the interaction of Dnmt3L, a partner of Dnmt3a and 3b, with chromatin (Ooi *et al*, 2007). Additionally there is evidence that the ADD domain of Dnmt3a and 3b that binds to unmethylated H3 is inhibited by H3K4me3 (Zhang *et al*, 2010b). My data suggests that a high CpG frequency and high G+C content might be enough in ES cells to protect against *de novo* DNA methylation.

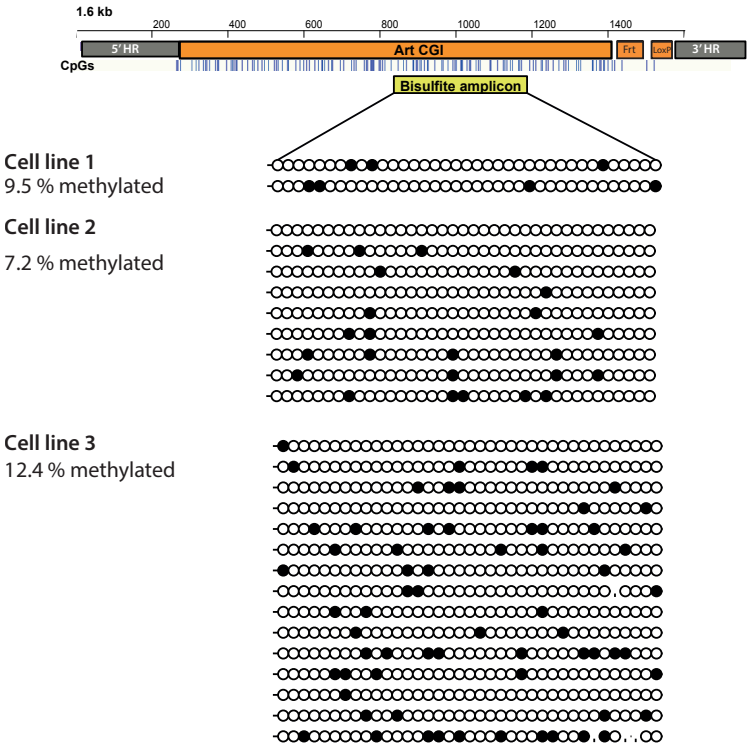


Figure 4.2.2-1 Artificial CGI in gene desert 2 remains unmethylated in mouse ES cells
Bisulfite sequencing of 3 cell lines containing the artificial CGI construct in gene desert 2. Scheme on top shows the inserted construct with the flanking gene desert arms and the remaining frt and LoxP sites. Scale is in bp. Blue vertical lines show position of CpGs. Sequenced amplicon is highlighted by yellow box. A methylated CpGs is depicted by a filled black circle, an unmethylated CpG by an empty white circle.

4.2.3. Differentiation of mouse embryonic stem cells into neuronal precursors leads to loss of H3K4me3 activity and gain of Polycomb at artificial CGI in a gene desert

As described above the artificial CGI inserted in a human gene desert created a bivalent chromatin state in mouse ES cells. A classic view of the characteristic of bivalent domains is that they keep key developmental genes in pluripotent cells poised for activation. Upon differentiation the repressive state can either be overcome by strong, activating stimuli or stay stably repressed in absence of transcriptional activation. Since the artificial CGI is comprised of a promoter-less CpG and G+C rich sequence in a gene desert and no transcriptional stimuli are expected we anticipated that repressive mechanism would take over. The 3 mouse ES cell lines containing the artificial CGI were differentiated into neural precursor cells (NPCs). Figure 4.2.3-1panel A shows the differentiation protocol used. When performing ChIP with antibodies against H3K4me3 from undifferentiated mouse ES cells and NPC it became apparent that the H3K4me3 levels were reduced in the NPCs at the CGI-like sequence in cell lines 1 and 2 (Figure 4.2.3-1 B). This effect could not be observed for cell line 3. However, when one compares H3K4me3 levels of the bivalent control genes it becomes apparent that in cell lines 1 and 2 the levels of K4me3 are reduced or the same in NPCs compared to ES cells. In cell line 3, however, this effect seems to be reversed. It shows higher levels of K4m3 at the control genes *HoxC8* and *HoxA9* in NPCs cells compared to ES cells. Additionally the H3K4me3 levels at the CGI are more than 5 fold higher than at the control genes, which is not the case for cell line 1 and 2. Therefore upon differentiation in cell line 3 a greater reduction of H3K4me3 needs to be achieved compared to the other 2 cell lines. When looking at the H3K27me3 ChIP in ES cells versus NPCs (Figure 4.2.3-2) it seems that the H3K27me3 over the inserted CGI-like sequence remain high in cell line 1 and to a lesser degree in cell line 2. Again the control levels in cell line 3 in ES cells are different, as levels at the bivalent control genes are not significantly higher than over the negative control region m15. In summary, there seems to be a reduction of H3K4me3 levels at the artificial CGI in NPCs compared to ES cells whereas H3K27me3 levels remain high. However, these results are very preliminary as the differentiation experiment was only performed once with every cell line. Clearly more biological replicates are needed in order to draw conclusions on the fate of chromatin over an inserted promoterless CGI. Additionally it is necessary to check for proper differentiation by analysing appearance of neural specific markers and decrease of pluripotency gene expression. The observed mixed results could also indicate that differentiating ES cells until

NPCs was not enough and that further differentiation into mature neurons would be required. Moreover, it would be interesting to determine the DNA methylation state of the artificial CGI in NPCs in comparison with ES cells to ask whether the unmethylated state can be perpetuated in differentiated cells or if the CGI-like sequence becomes methylated in absence of transcription.

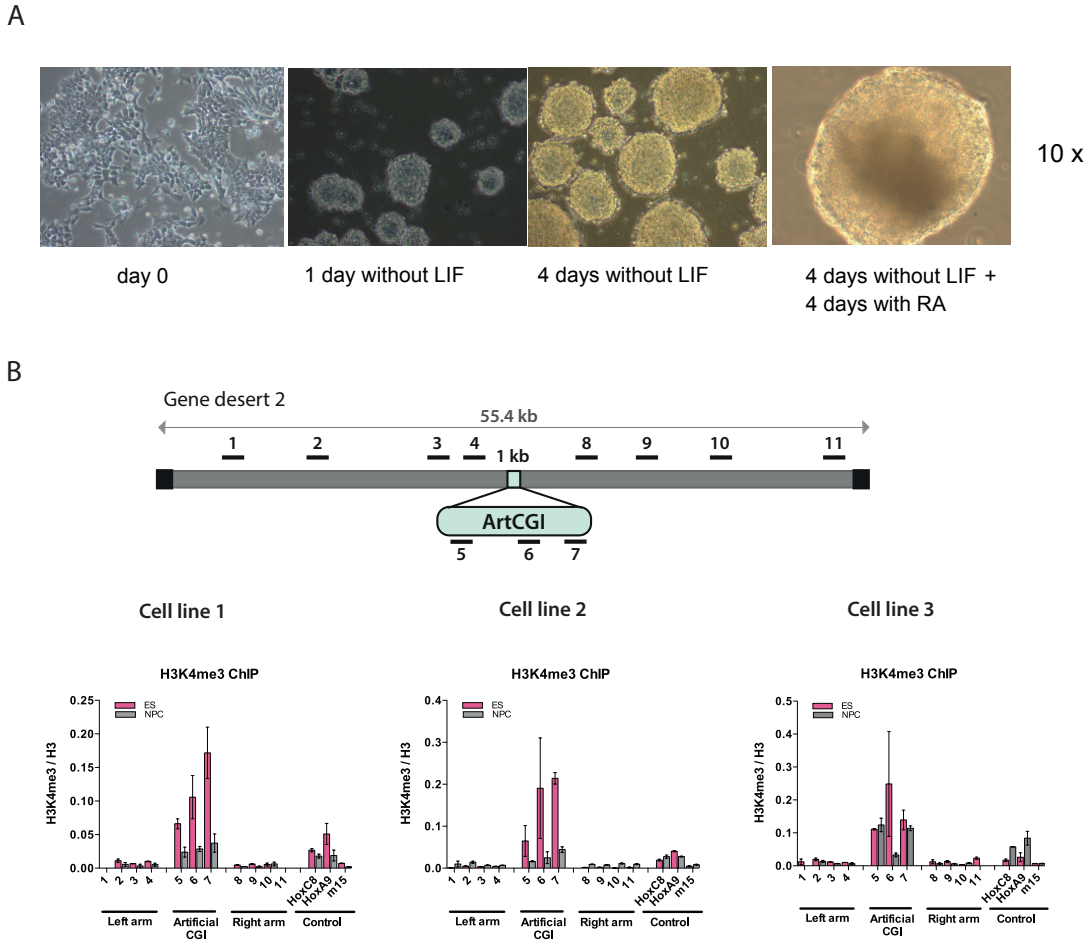


Figure 4.2.3-1 Reduced levels of H3K4me3 over artificial CGI in neural precursor cells
A: Undifferentiated mouse ES cells were cultured for 4 days without leukemia inhibitory factor (LIF) and then another 4 days in the presence of retinoic acid (RA), all panels show 10x magnification. B: Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars indicate position of primers used for Q-PCR (length of amplicons not drawn to scale): H3K4me3 ChIP for 3 cell lines in ES cells versus neural precursor cells. Y-axis: % of Input of H3K4me3 over % of Input of pan-H3 antibody. Controls: TSS of bivalent genes HoxC8 and HoxA9; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates

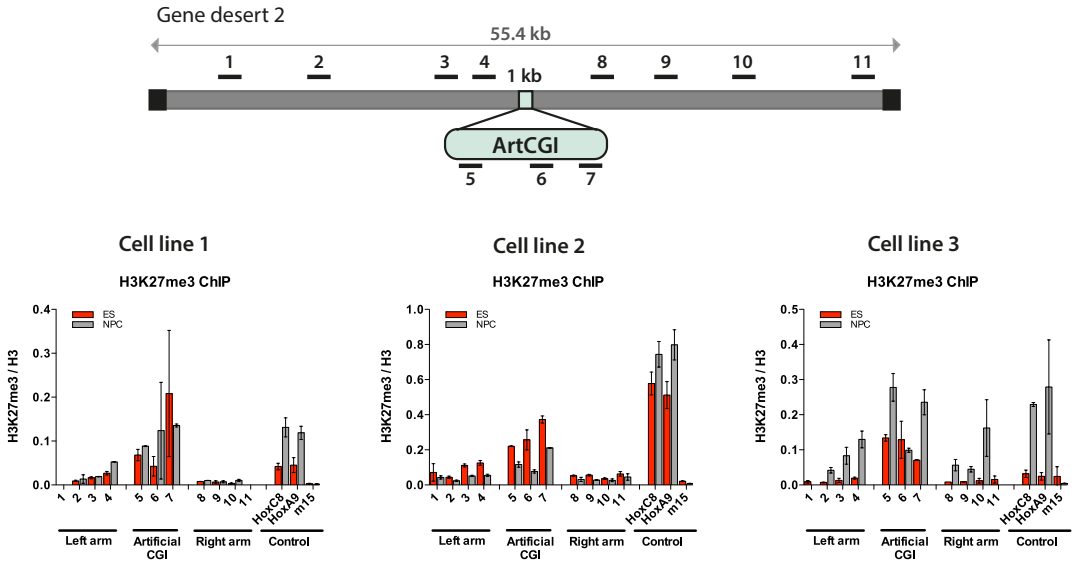


Figure 4.2.3-2 H3K27me3 levels over artificial CGI remain high in neural precursor cells
Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars indicate position of primers used for Q-PCR (length of amplicons not drawn to scale): H3K27me3 ChIP for 3 cell lines in ES cells versus neural precursor cells. Y-axis: % of Input of H3K4me3 over % of Input of pan-H3 antibody. Controls: TSS of bivalent genes HoxC8 and HoxA9; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates.

4.2.4. Cfp1 is detected at artificial CGI in gene desert

We have shown that an artificial CGI-like sequence establishes H3K4me3 and H3K27me3 histone marks. It was also shown that Suz12, a component of PRC2, is present, strengthening the notion that a bivalent domain is created. In case of H3K4me3, testing for the presence of the corresponding histone methyltransferase was not as straight forward. We attempted to perform a ChIP with antibodies against Cfp1. Since the signal of H3K4me3 at the CGI is much lower than that detected at TSS and the ChIP efficiency of the α-Cfp1-antibody is significantly lower than that of an α-H3K4me3 antibody we suspected that our α-Cfp1 antibody did not provide enough sensitivity to detect this protein at the integrated CGI.

As an alternative we obtained a transgenic *Cfp1*-GFP tagged mouse ES cell line from Francis Stewart's laboratory. This line can then be used to integrate the artificial CGI within a gene desert and ask by performing an anti-GFP ChIP if Cfp1 is present at the inserted CGI. The cell line was generated by BAC transgenesis of a *Cfp1*-GFP BAC construct. This BAC contains the whole *Cfp1* gene with all the regulatory elements and GFP fused to the last

codon of *Cfp1* in order to create a C-terminal GFP tag. Since the *Cfp1*-GFP tagged ES cells already contain a neomycin selection cassette it was not possible to use the previously used BAC construct with the artificial CGI as this construct is also based on selection with neomycin. In order to avoid that problem the neomycin cassette was exchanged with a blasticidin-containing cassette flanked by Rox sites. This cassette exchange was performed by amplifying the blasticidin gene with PCR primers that contained 50 bp homology arms corresponding to the artificial CGI and the 3' homology arm from the original gene desert 2 plus ArtCGI BAC. The neomycin cassette was then exchanged with this amplified blasticidin construct by recombineering. Successfully recombined clones were identified by control digest and sequencing. The *Cfp1*-GFP tagged cell line was then transformed with the BAC containing the gene desert 2 with the artificial CGI-like and the blasticidin cassette in it's middle. As described before ES cell colonies were screened for the presence of the full length BAC. Three cell lines were expanded further and in a second targeting step the selection cassette was excised using a Dre recombinase that recognizes Rox sites (Anastassiadis *et al*, 2009). Southern blots were performed to identify the successful excision of the cassette (not shown) and all three lines were used for ChIP experiments. Initially ChIPs with antibodies against H3K4me3 and H3K27me3 were performed to show that the artificial CGI establishes a bivalent domain in those *Cfp1*-GFP tagged ES cells just in wt ES cells. Figure 4.2.4-1 shows that all three cell lines have H3K4me3 levels equal or a little higher than that of the TSS of the bivalent genes *HoxC8* and *HoxA9*. As before no signal was detected over the flanking gene desert region. As in the wild type ES cells a clear peak of H3K27me3 was observed over the inserted CGI spreading into the adjacent gene desert.

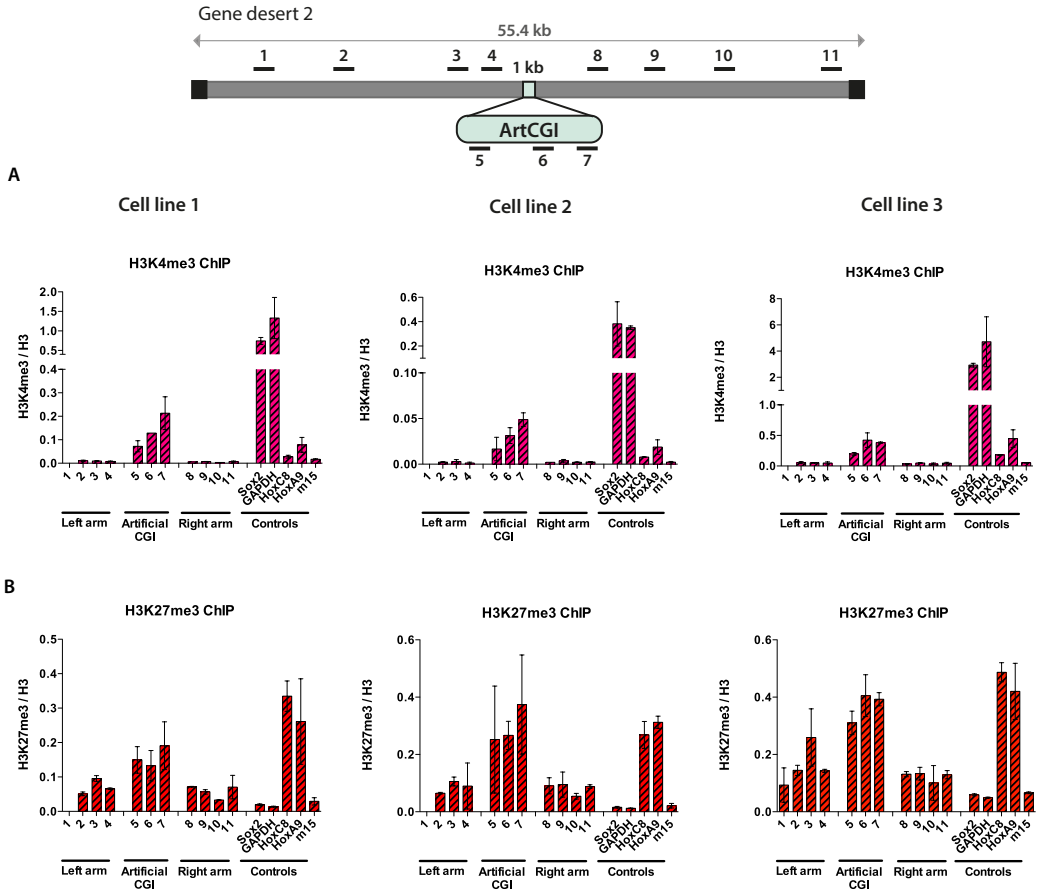


Figure 4.2.4-1 A bivalent domain is established at artificial CGI in Cfp1-GFP tagged ES cells
Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars indicate position of primers used for Q-PCR (length of amplicons not drawn to scale). A: H3K4me3, B: H3K27me3 ChIP for 3 cell lines. Y-axis: % of Input of H3K4me3 over % of Input of pan-H3 antibody. Controls: TSS of active genes Sox2 and GAPDH and of bivalent genes HoxC8 and HoxA9; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates

Once it was established that a bivalent domain was formed over the artificial CGI in *Cfp1*-GFP cells we wanted to know if Cfp1 is present as well. Thus a ChIP with an antibody against GFP was performed with three cell lines in two replicate experiments (except cell line 3). For these ChIPs beads already coupled to an anti-GFP antibody were used (GFP-TRAP system, see material and methods). As can be seen in Figure 4.2.4-2 there is a clear enrichment of Cfp1 over the TSS of the active genes *Sox2* and *GAPDH* in the *Cfp1*-GFP tagged cells in comparison to wt ES cells. In cell line 1 there is also clear difference in Cfp1 levels at active genes versus bivalent genes and at the negative control region m15. Both differences are much less pronounced in cell lines 2 and 3. When looking at Cfp1 signal over the artificial CGI in the 1st X-link experiment it becomes apparent that in all three cell lines there is a clear enrichment in *Cfp1*-GFP tagged cells in contrast to wt ES cells containing the

artificial CGI. The percentage of input that was precipitated over the island is in all lines similar (around 0.1%) despite the fact that the enrichment over control genes was less in cell line 2 and 3. A second X-link for cell line 1 and 2 shows a strong signal over the CGI that is not present in wt cells. The enrichment seems considerably higher than over the active control genes, which is surprising as it is expected that Cfp1 levels at active genes are higher than at bivalent regions. This was not the case for the H3K4me3 ChIPs where K4me3 levels were consistently higher at active genes than over the inserted CGI. It is not clear why in this second experiment there is a discrepancy between levels of H3K4me3 and Cfp1 in comparison to active genes. Despite this variation it is seems that Cfp1 is present at the inserted artificial CGI specifically and not at the adjacent gene desert region. This confirms previous results that a CpG rich sequence attracts the histone methyltransferase SET1A/B presumably via the CxxC domain of Cfp1 (Thomson *et al*, 2010).

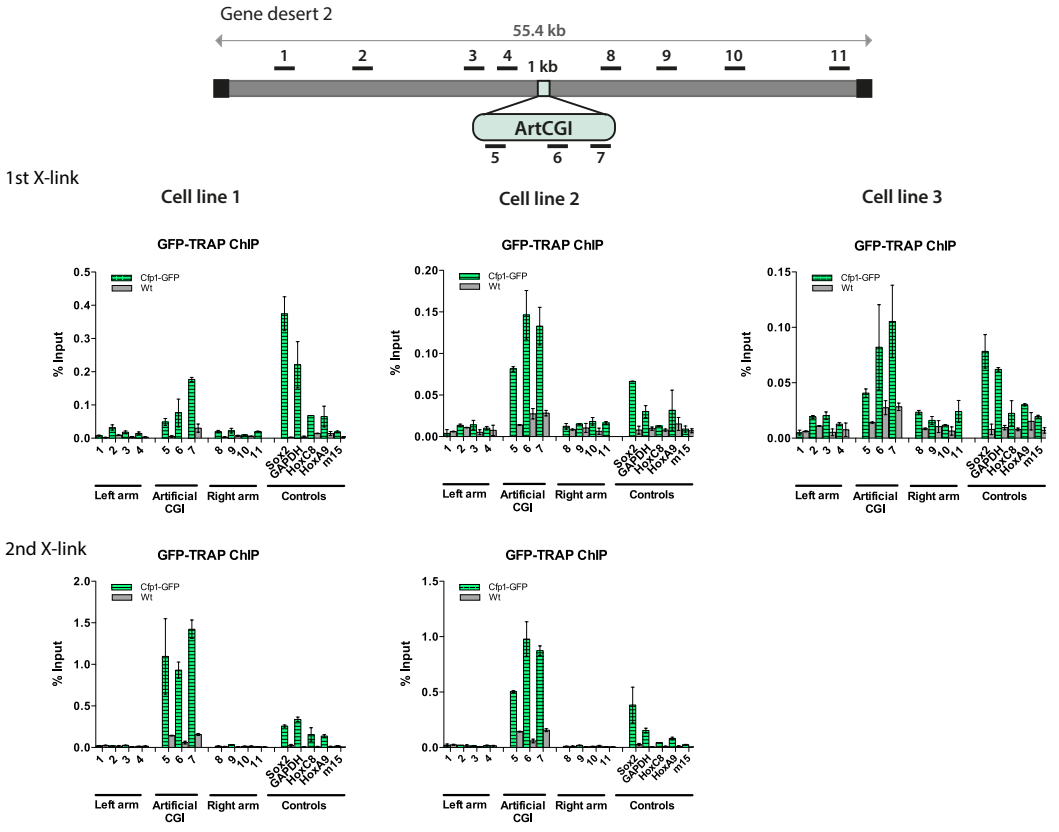


Figure 4.2.4-2 GFP Cfp1 is detected at artificial CGI in gene desert 2
Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars indicate position of primers used for Q-PCR (length of amplicons not drawn to scale). Anti GFP-ChIPs with beads already coupled to GFP antibody (GFP-TRAP beads) is shown for 3 cell lines. Y-axis: % of Input. Controls: TSS of active genes Sox2 and GAPDH and of bivalent genes HoxC8 and HoxA9; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates

4.2.5. Cfp1 is not required for the formation of H3K4me3 at CGI in gene desert

In order to investigate if the formation of a bivalent domain at the inserted artificial CGI-like sequence is dependent on Cfp1 the artificial CGI was introduced into *Cfp1* -/- mouse ES cells. *Cfp1* null mouse ES cells were created in the laboratory of David Skalnik (Carlone & Skalnik, 2001). As this deletion proved to be embryonic lethal no viable *Cfp1* null mice were obtained. It was, however, possible to acquire mouse ES cells from blastocysts (Carlone *et al*, 2005). Since those *Cfp1*-/- cells were already resistant to neomycin it was again necessary to transform them with a BAC construct containing a blasticidin selection cassette (described in 4.2.4). Again three independent cell lines with low copy number integration were chosen and after excision of the selection cassette they were analysed by ChIP. It was also confirmed that the *Cfp1* gene was really deleted by performing PCRs across this region. Figure 4.2.5-1 shows the result of H3K4me3 and H3K27me3 ChIPs in all 3 cell lines. As previously reported H3K4me3 levels were strongly reduced at CGIs of strongly expressed genes (Clouaire *et al*, 2012). This reduction can be seen in panel A where the H3K4me3 signal at the actively transcribed genes *Sox2* and GAPDH was similar to that of the bivalent genes HoxC8 and HoxA9. This was not the case in wt ES cells where the H3K4me3 levels at active genes were much higher (for example see Figure 4.2.1-4). Concerning the H3K4me3 levels over the ArtCGI it is obvious that an H3K4me3 domain was established at the inserted CGI-like sequence even in the absence of Cpf1. Again the strength of the signal was similar or slightly higher than that of the bivalent control genes. Also H3K27me3 is established over the CGI in *Cfp1* deficient cells in the same manner as observed in *Cfp1*-wt cells. This is expected as the PRC2 complex writing the H3K27me3 mark should not be altered in *Cfp1* deficient cells (Figure 4.2.5-1panel B). In summary while Cfp1 may contribute to the formation of an H3K4me3 peak at the artificial CGI, its presence is not necessary.

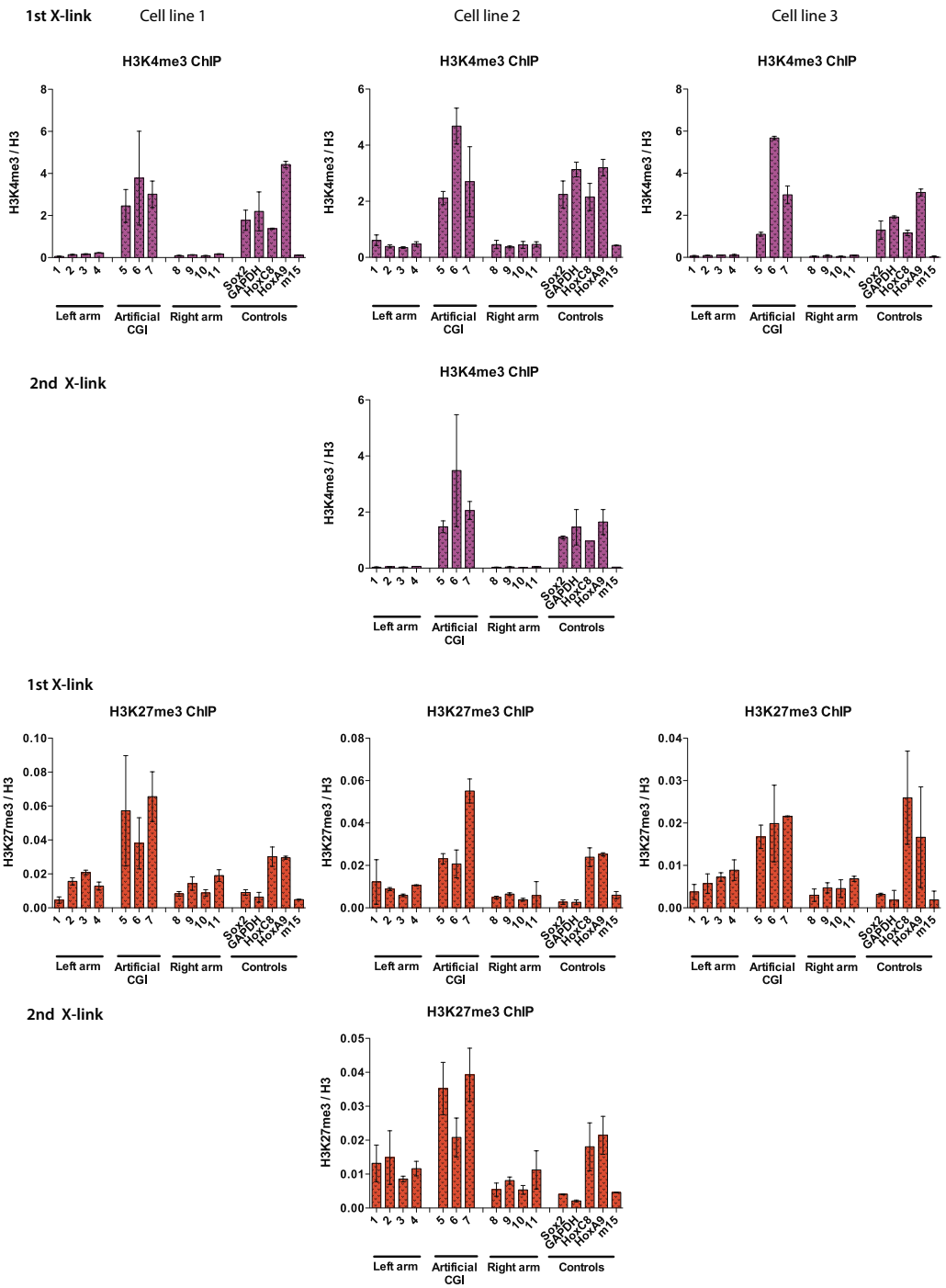


Figure 4.2.5-1 Cfp1 is not required for H3K4me3 establishment over artificial CGI in gene desert 2

Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars indicate position of primers used for Q-PCR (length of amplicons not drawn to scale). A: H3K4me3 ChIP for 3 cell lines. B: H3K27me3 ChIP for 3 cell lines. Y-axis: % of Input of H3K4me3 or H3K27me3 over % of Input of pan-H3 antibody. Controls: TSS of active genes Sox2 and GAPDH and of bivalent genes HoxC8 and HoxA9; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates

4.3. Discussion

Unlike with the PuroGFP construct, 3 different cell lines were obtained containing the artificial CGI thanks to an improved Flp transformation protocol. Having different cell lines available for analysis offers the possibility to examine potential effects of different integration sites. All three cell lines show the same histone modification patterns, presence of Suz12 and absence of RNA polymerase II, suggesting that the insulator function gained by the human gene desert BAC is sufficient to protect against positional effects. One concern of adopting the strategy presented here for integrating CGI-like sequences in a gene desert was that integration of multiple copies might influence the results. In order to circumvent this issue, quantitative PCR was employed to analyse copy number integration. Ideally only cell lines with a single copy would have been used for this study, which proved to be challenging. Therefore clones with a low copy number (1, 1-2 and 4) were taken forward (Figure 4.2.1-2). When levels of different histone modifications or proteins over the CGI are compared between the different cell lines it seems that the slight differences in determined copy numbers do not influence the result. Cell line 3 proved to have the highest number of integrations but this is not reflected in higher levels of any of the analysed modifications in comparison with the other two cell lines. In summary we conclude that the chosen approach is suitable to analyse the influence of base composition on chromatin establishment.

4.3.1. Is the bivalent domain observed at the artificial CGI truly bivalent?

The results presented above indicate that a CpG rich sequence with an overall high G+C content integrated in a gene desert region is able to recruit histone-modifying enzymes to create a bivalent domain. This is in agreement with earlier findings that G+C rich DNA can recruit the Set1 complex via Cfp1 and create a novel domain of H3K4me3 (Thomson *et al*, 2010) and polycomb group proteins to establish H3K27me3 (Mendenhall *et al*, 2010; Lynch *et al*, 2011).

Although ChIP experiments are highly informative regarding the genomic localization and the correlation of different marks can be assessed in cell populations, they don't provide information about physical co-existence of modifications on the same nucleosome. Therefore it cannot be ruled out that the appearance of a bivalent domain at the artificial CGI constitutes an average cell population where the CGI in some cells is marked by H3K4me3 and in others by H3K27me3. Indeed, Hong and co-workers, for example, found that by fractionating human ES cells into different subpopulations according to cell specific

markers, apparent bivalent domains can be resolved into monovalent signatures (Hong *et al*, 2011). Another study claims that the occurrence of bivalent domains is largely due to ES culture medium containing foetal bovine serum (Marks *et al*, 2012). They show that culturing ES cells in a serum free medium, the so-called 2i (inhibitors of two kinases Mek and GSK3) medium leads to changes in H3K27me3 deposition whereas the pattern of H3K4me3 is similar between the different culturing conditions. Cells treated with 2i medium exhibit reduced prevalence of the repressive histone modification H3K27me3 at promoters and higher levels at satellite regions, while the global levels stay the same. Only around 1000 bivalent domains were detected, compared to more than 3000 in serum-cultured ES cells (Marks *et al*, 2012). In the future it would be interesting to assess if the bivalent domain detected at the artificial CGI is caused by culturing cells in serum containing medium.

Sequential ChIP is a method that resolves the issue about cell average bias and provides information about the chromatin state on mononucleosomes. It relies on a first IP with, for example, an antibody against H3K4me3 followed by a second IP with an antibody against H3K27me3. In some studies sequential ChIPs were used in order to address the question of whether bivalent domains exist within one cell on the same nucleosome. Bivalent domains were demonstrated to exist in zebrafish (Vastenhouw *et al*, 2010) and in human (Pan *et al*, 2007) and mouse embryonic stem cells (Voigt *et al*, 2012) but were not detected in *Xenopus* or *Drosophila* embryos (Akkers *et al*, 2009; Schuettengruber & Cavalli, 2009). However, it needs to be taken into account that results obtained by re-ChIP experiments might be biased by insufficient antibody specificity, contamination with initial antibody or oligonucleosomes. If oligonucleosomes are present in the material used for ChIP the possibility remains that the marks identified are actually present on neighbouring nucleosomes. It was shown that H3K4me3 and H3K27me3 cannot co-occur on the same H3 tail (Young *et al*, 2009) but as each mononucleosome contains two copies of each histone it is conceivable that one H3 copy could be marked by K4me3 whereas the other one is marked by K27me3. A recent report has asked if asymmetrically modified nucleosomes occur by antibody based affinity purification of mononucleosomes followed by mass spectrometry analysis (Voigt *et al*, 2012). The authors found that H3K4me3 and H3K27me3 coexist *in vivo* on the same nucleosome but on opposite H3 tails.

In the future it would be interesting to perform sequential ChIP analysis on the cell lines carrying the artificial CGI in a gene desert in order to address the question if both H3K4me3

and H3K27me3 marks occur at the same nucleosome or if in fact the observed bivalency is a result of cell population average.

4.3.2. CpG islands and enhancers

In Figure 4.2.1-6 it was shown that low levels of H3K4me1, a mark correlated with enhancers, were established over the artificial CGI. However, it is not clear what the apparent co-occurrence of H3K4me1 and H3K4me3 at the artificial CGI means. Do these two marks occur at the same nucleosome or is this an effect of population average and there are some cells that have H3K4me1 and others that show H3K4me3? Alternatively H3K4me1 might be a transient state on the way to K4me3.

Enhancers are thought to activate transcription by delivering important accessory factors to promoters in order to enhance either the formation of the pre-initiation complex or the transition from initiation to elongation (Calo & Wysocka, 2013). H3K4me1 and H3K27ac are the major histone modification types found adjacent to H2A.Z//H3.3 containing nucleosomes, which are histone variants commonly found at enhancers (Jin & Felsenfeld, 2007). During enhancer activation H3K4me1 presence precedes nucleosome rearrangement and H3K27ac formation, indicating that this chromatin mark might pre-mark enhancer regions before they actually become fully activated. H3K4me1 might be established at ubiquitous genomic loci that allow recruitment of activating histone-modifying enzymes. This mark could then offer a window of opportunity within which enhancer activation can occur (Calo & Wysocka, 2013). One possible function of H3K4me1 could be to keep distal regulatory regions from being targeted for *de novo* DNA methylation. The fact that the artificial CGI-like sequence analysed in this study showed low levels of H3K4me1 and absence of methylation supports this notion.

It might seem surprising that enhancers are not globally marked by H3K4me3, given the fact that one of the pathways to establish this mark is via interaction with the serine 5 phosphorylated form of RNA PolII, which is enriched at most enhancers (Bonn *et al*, 2012). A recent study offered an explanation by showing that deletion of Cfp1, a member of the Set1 histone methyltransferase complex, results in a depletion of H3K4me3 from active CGI containing promoters and in the ectopic appearance of H3K4m3 at distal regulatory elements such as enhancers (Clouaire *et al*, 2012). Rescue experiments with a mutant that is deficient in the CpG binding domain CxxC does not prevent the aberrant accumulation of H3K4me3 at enhancers but re-establishes H3K4me3 levels at promoters (Clouaire *et al*, 2012). This

suggests that one main function of Cfp1 is to restrict recruitment of H3K4me3 histone methyltransferases to CpG rich regions.

Interestingly, there exists a class of enhancers that are enriched for H3K4me1, but lacking H3K27ac, are also marked by H3K27me3 and bound by the Polycomb complex PRC2 (Calo & Wysocka, 2013). These so called "poised enhancers" are located near poised or bivalent promoters, characterized by the simultaneous presence of H3K4me3 and H3K27me3 and also association with PRC2. Taken together, the fact that H3K4me1 and me3 as well as H3K27me3 marks were found at the artificial CGI could mean that a CpG rich / G+C rich sequences recruit histone-modifying enzymes by default setting the stage towards a more accessible chromatin waiting for further instructions by, for example, transcription factors.

4.3.3. Is a high CpG density and high G+C content enough to be kept free of DNA methylation?

In the present study it was shown that artificial CGI-like sequences introduced into human gene desert 2 did not become methylated (Figure 4.2.2-1). Despite many years of research the exact rules that govern DNA methylation establishment are still not known. Initially it has been proposed that transcription is required to prevent a DNA sequence from becoming methylated, as it provokes high levels of H3K4me3 methylation (Takeshima *et al*, 2009; Brandeis *et al*, 1994; Macleod *et al*, 1994). However, despite the presence of H3K4me3, no form of RNA PolII was detected over the inserted CGI-like sequence indicating that transcription is not always necessary for keeping a region methylation free.

This is in accordance with a recent study by Lienert at al, where around 50 different sequences were inserted into a transcriptionally inert locus. These sequences autonomously recapitulated correct DNA methylation in absence of transcription (Lienert *et al*, 2011). However these authors argue against CpG density alone as the parameter that keeps a sequence methylation free, as they found that 7 out of 10 sequences from the *E.coli* genome that had an average length of 780 bp and varied in CpG density from 4.4 to 6.8 CpGs per 100 bp, became methylated. This CpG frequency though is on the lower end of what normally constitutes a CGI. The 3 fragments that did not become methylated were among those with the higher CpG frequency, although some other fragments with the same CpG density did become methylated. In those cases where the fragments stayed unmethylated, the active chromatin mark H3K4me2 was detected suggesting that the modification state of the H3 tail plays a role, which is in accordance with work implicating H3K4 methylation in

inhibiting *de novo* DNA methylation (Ooi *et al*, 2007). In agreement with this, the artificial CGI-like sequence tested in the present study displayed presence of H3K4me3 and absence of DNA methylation. They suggest that so called methylation determining regions (MDRs) are responsible for conferring a certain methylation pattern and the absence of these regions in *E. coli* sequences is responsible for their susceptibility to DNA methylation even though they contain a high CpG density (Lienert *et al*, 2011). Another study supports this view by showing that the maintenance of unmethylated states is not dependent on "CpGness" but on the presence of sequence specific motives (Straussman *et al*, 2009). Additionally, cooperative binding of the transcription factors Sp1, Nrf-1, and YY1 in normal monocytes correlates with protection from CGI methylation in leukaemia cells (Gebhard *et al*, 2010). It is likely that a combination of transcription factor binding sites, attraction of histone modifying enzymes via CpG density and transcription are responsible for establishing a specific DNA methylation pattern.

4.3.4. Are other histone methyltransferases compensating for the absence of Cfp1?

In comparison to yeast, which has only one H3K4 methylase, mammals carry at least six H3K4 methyltransferases: Set1A, Set1B, MLL1, MLL2. MLL3 and MLL4, which are all part of complexes that are related to the yeast Set1 complex (COMPASS) (Eissenberg & Shilatifard, 2010). Interestingly, deletion of any one of the MLL family members has only minimal effects on global H3K4me3 levels, suggesting redundancy among the MLL complexes (Wang *et al*, 2009). In contrast, deletion of WDR5, RbBP5, and Ash2L, integral shared core subunits that are necessary for the methylation activity of these complexes, greatly reduces H3K4 methylation levels (Dou *et al*, 2006).

The finding here that a novel mark of H3K4me3 is created over the artificial CGI in absence of Cfp1, which is a subunit of both the Set1A and B histone methyltransferase complexes, suggests that there might be compensating activities present. Two candidates MLL1 and MLL2, H3K4me3 methyltransferases that contain CxxC domains, are expressed in mouse ES cells (Glaser *et al*, 2006; Jiang *et al*, 2011). Consistently, mouse ES cells are viable in the absence of Cfp1 but their failure to differentiate implicates a role for Cfp1 in lineage commitment (Carlone *et al*, 2005). Moreover, loss of Cfp1 in mice results in early embryonic lethality (Carlone & Skalnik, 2001). There have been reports that implicate MLL proteins in roles during early development, whereas Set1 becomes the dominant H3K4 methyltransferase at later developmental stages (Ardehali *et al*, 2011; Mohan *et al*, 2011)

(Wu *et al*, 2008). This indicates that the contribution of CxxC proteins to chromatin establishment changes during differentiation. Contrary to depletion of Cfp1, which is only present in Set1 complexes and results in a decrease of H3K4me3 at active genes without affecting non-productive genes (Clouaire *et al*, 2012) , depletion of subunits shared by Set1 and MLL1/2 complexes leads to a more pronounced reduction in H3K4me3 at active and poised genes (Ang *et al*, 2011; Jiang *et al*, 2011). The chromatin at the inserted artificial CGI-like sequence resembles more poised genes with respect to medium H3K4me3 levels, no RNA polymerase II and the presence of H3K27me3. Therefore, it is possible that H3K4me3 at artificial CGI insertions depends on other CxxC domain proteins, such as MLL1 and/or MLL2 even though Cfp1 was shown to be present at the artificial CGI. Indeed, very recently it was shown that MLL2 is the H3K4 methyltransferase in mammals that is responsible for trimethylation of H3K4 at bivalent genes (Hu *et al*, 2013).

5. High CpG frequency is sufficient to establish a bivalent domain in gene desert region

5.1. Introduction

As discussed in chapter 4 an artificial CGI-like sequence is enough to establish a novel peak of H3K4me3 and H3K27me3. It is less clear, however, what features exactly are responsible for creating those chromatin marks. Is it the frequency of CpGs, is it the G+C content, is it both and are the requirements different for H3K4me3 and H3K27me3? In the case of the establishment of H3K4me3 one hypothesis is that the high density of CpG in CGIs is crucial as it attracts H3K4 methyltransferases. This could be either the Set1 complex via the CxxC domain of Cfp1 or other CxxC domain containing methyltransferases such as MLL1 or 2. But is a high CpG frequency sufficient or does the overall G+C content play a role as well? How Polycomb proteins are recruited to their target genes has not been definitely resolved either. It is known that components of the PRC2 complex, which contains the enzyme responsible for methylating H3K27, correlate strongly with CGIs. To date though it is not clear how these proteins are recruited to CGIs. So far no CxxC containing proteins have been shown to perform this function. Also it is unclear if CpGs are indeed the responsible factor. A recent report showed that a GC rich sequence from *E. coli* is sufficient to recruit Polycomb proteins (Mendenhall *et al*, 2010). However, in this study a distinction between CpGs and G+C content was not made and it remains unknown what feature of CGIs attract Polycomb proteins.

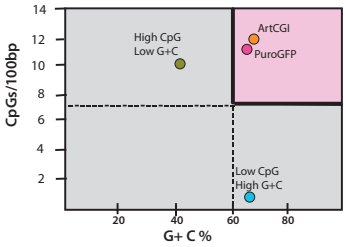
5.2. Results

5.2.1. A high G+C content is not sufficient for creation of bivalent domain

In order to answer those questions two additional CGI like sequences with perturbed base composition were created. One sequence has the same length of 1000bp as the average CGI and contains a G+C content of 64.4%, typical of CGIs. In contrast the number of CpGs is reduced to 10, which corresponds to an observed-over-expected CpG ratio of 0.1, a value typical for the bulk genome. In addition to these base composition requirements, Sp1 binding sites were avoided, but otherwise the sequence was designed randomly. This sequence will be referred to as Low CpG / High G+C (Lo/Hi) and its exact sequence can be found in the

appendix. Figure 5.2.1-1 shows an overview of this Low CpG / High G+C sequence in comparison with the base composition of the PuroGFP and the first artificial CGI described in the previous chapters.

A



B

	Length in bp	G+C content	A+T content	Nr of CpGs	o/e CpG ratio	CpGs/ 100 bps	SpI sites
PuroGFP	1267 bp	66.5%	33.5%	140 CpGs	1,0	11	1
Artificial CGI	1055 bp	69.5%	30.5%	127 CpGs	1.0	12	-
High CpG/ Low G+C	1007 bp	40.3%	59.7%	100 CpGs	2.5	10	-
Low CpG/ High G+C	1008 bp	64.4%	35.6%	10 CpGs	0.1	1	-

C

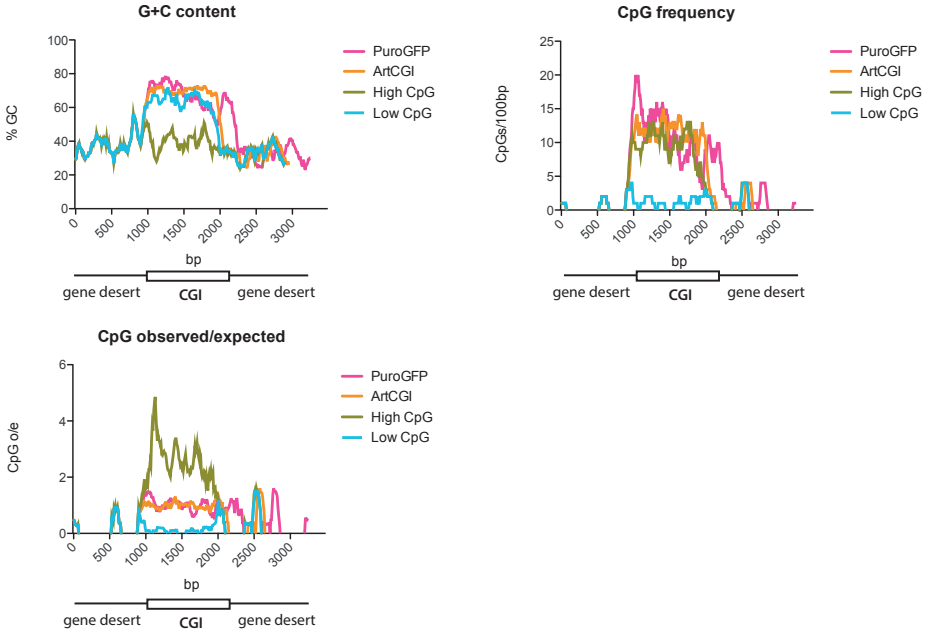


Figure 5.2.1-1 Overview of CGI like sequences used in this study
A: Colored dots represent the different sequences used in this study. Their position depends on CpG frequency and G+C content. Pink box indicates parameters typically for CGIs. B: Table showing length and base composition of constructs. O/e = observed over expected CpG ratio. Note that the formula calculating the o/e ratio (Nr of CpGs /(Nr of C × Nr of G)) × Nr of nucleotides in the sequence) takes the overall G+C content into account. Therefore the High CpG/ Low G+C sequence has an unusually high o/e ratio. C: Different constructs are blotted depending on % of G+C, CpGs per 100 bp or CpGs observed over expected. X-axis length in bps of CGI like construct flanked either side by 1 kb gene desert.

The Low CpG / High G+C sequence was introduced into gene desert 2 as described in Chapter 4. ES cells were transfected with the construct and screened for low copy number integration (Figure 5.2.1-2). The three cell lines marked with a red box showed a copy number of around one and were therefore chosen for further use. The selection cassette was

excised using Flp and positive clones for each cell line were identified by PCR and Southern blot, expanded and analysed by ChIP.

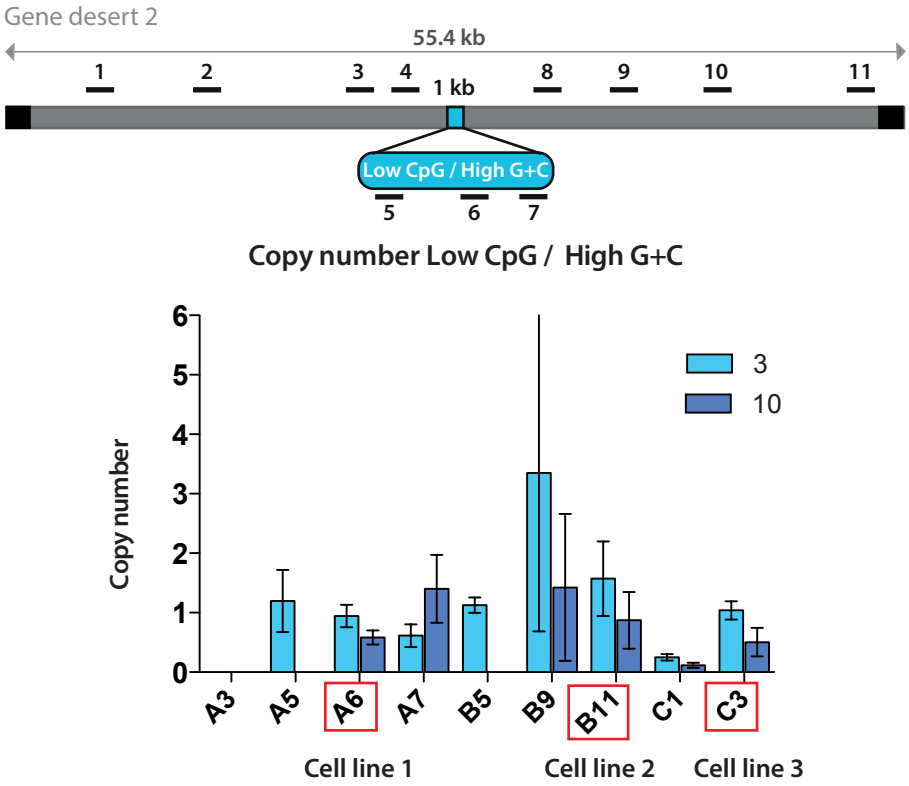


Figure 5.2.1-2 Copy number analysis of Low CpG / High G+C sequence in gene desert
Mouse ES cell clones with integrated Low CpG / High G+C construct were analysed by Q-PCR for copy number integration. 3 and 10 respectively refers to the primer pair used in the PCR. Values normalized to that of Sox2 = 2 copies per cell. Red boxes indicate chosen cell lines. Error bars indicate standard deviation of PCR triplicates. Scheme above indicates construct inserted in gene desert 2 and position of primers (not drawn to scale).

Figure 5.2.1-3 shows 2 independent biological replicates of ChIP with anti H3K4me3 antibodies in all three cell lines. Strikingly, the H3K4me3 signal was absent completely over the inserted Lo/Hi sequence for cell lines 1 and 2 with the signal remaining high as expected over the active positive control genes *GAPDH* and *Sox2*. This result suggests that in order to create a novel H3K4me3 mark a high frequency of CpGs is necessary and that an overall high G+C content is not sufficient for establishment of this mark, strengthening the notion that CpGs are important for attracting histone-modifying enzymes via their CxxC domain. However, in cell line 3, there is a low but significant amount of H3K4me3 detected over the Lo/Hi insertion.

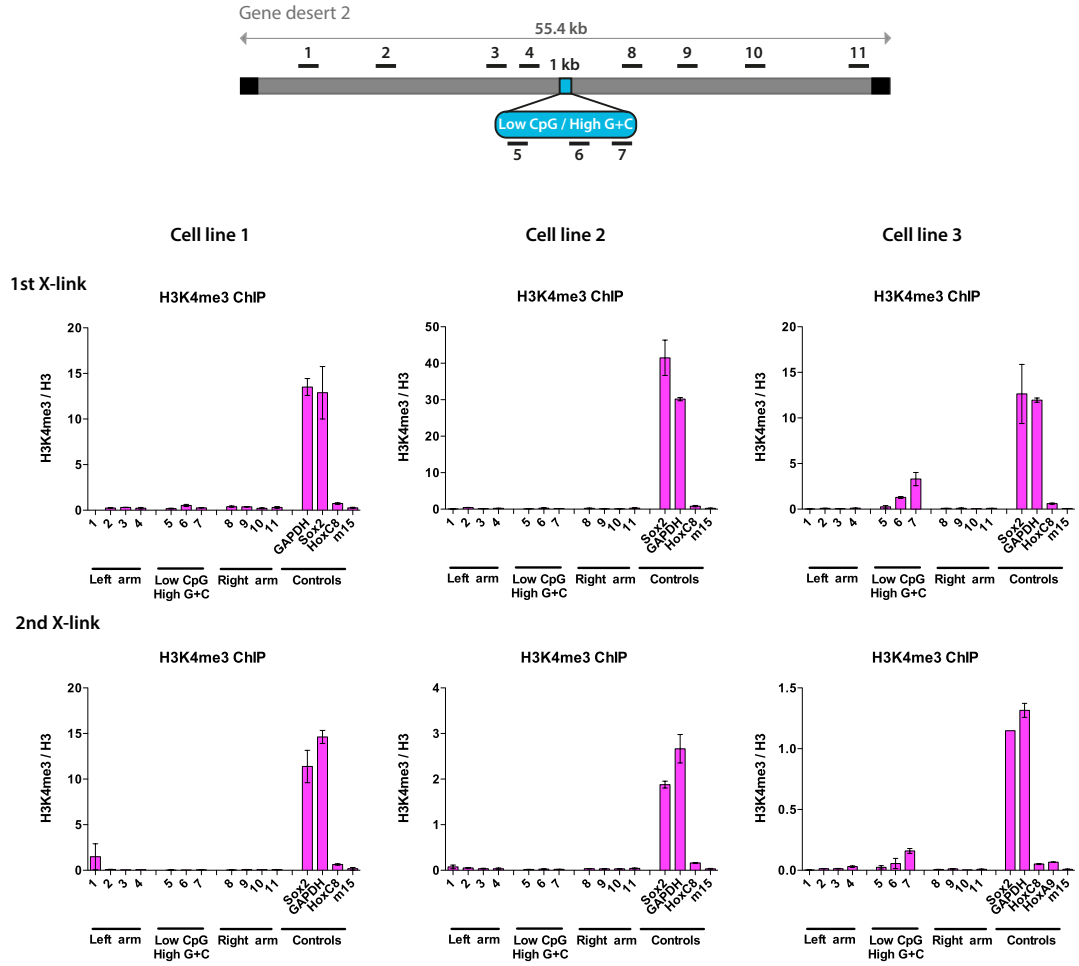


Figure 5.2.1-3 Is a high G+C content sufficient to esablish a novel H3K4me3 peak over the Low CpG / High G+C construct in gene desert 2?
H3K3me3 ChIP, 2 independent X-links for 3 cell lines. Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars indicate position of primers used for Q-PCR (length of amplicons not drawn to scale). Y-axis: % of Input of H3K4me3 over % of Input of pan-H3 antibody. Controls: TSS of active genes Sox2 and GAPDH, of bivalent gene HoxC8; m15= negative control region on mouse chromosome 15, indicates background levels. Error bars indicate standard deviation of PCR replicates.

A possible explanation for the presence of H3K4me3 observed in the third cell line would be failed excision of the selection cassette. If the neomycin indeed was not excised properly this could mean that the neomycin gene adjacent to the inserted Lo/Hi sequence is still present. As this gene was active during the selection process it is expected that high levels of H3K4me3 mark the promoter region of neomycin, which could influence the chromatin state of the adjacent Lo/Hi sequence. In order to test this hypothesis, the third cell line (Cell line 3 clone E10) was reanalysed and as can be seen in Figure 5.2.1-4 the selection cassette was indeed not excised completely. In contrast other clones from this cell line 3, E12 and G9

(same integration of BAC but different clone after transfection with Flp) did not show any traces of the cassette. When a ChIP was performed with clone E12 it became apparent that indeed our hypothesis was correct as the H3K4me3 signal was totally abolished over the integrated Lo/Hi sequence. This result confirms that it is essential to go through the process of excising the selection cassette. The presence of an active gene nearby can influence the establishment of chromatin at the inserted sequences, which would impede the interpretation of sequence influence on chromatin. In summary it can be concluded that a sequence with low frequency of CpGs but high G+C content is not sufficient of H3K4me3 establishment, emphasizing the importance of the CpG dinucleotide.

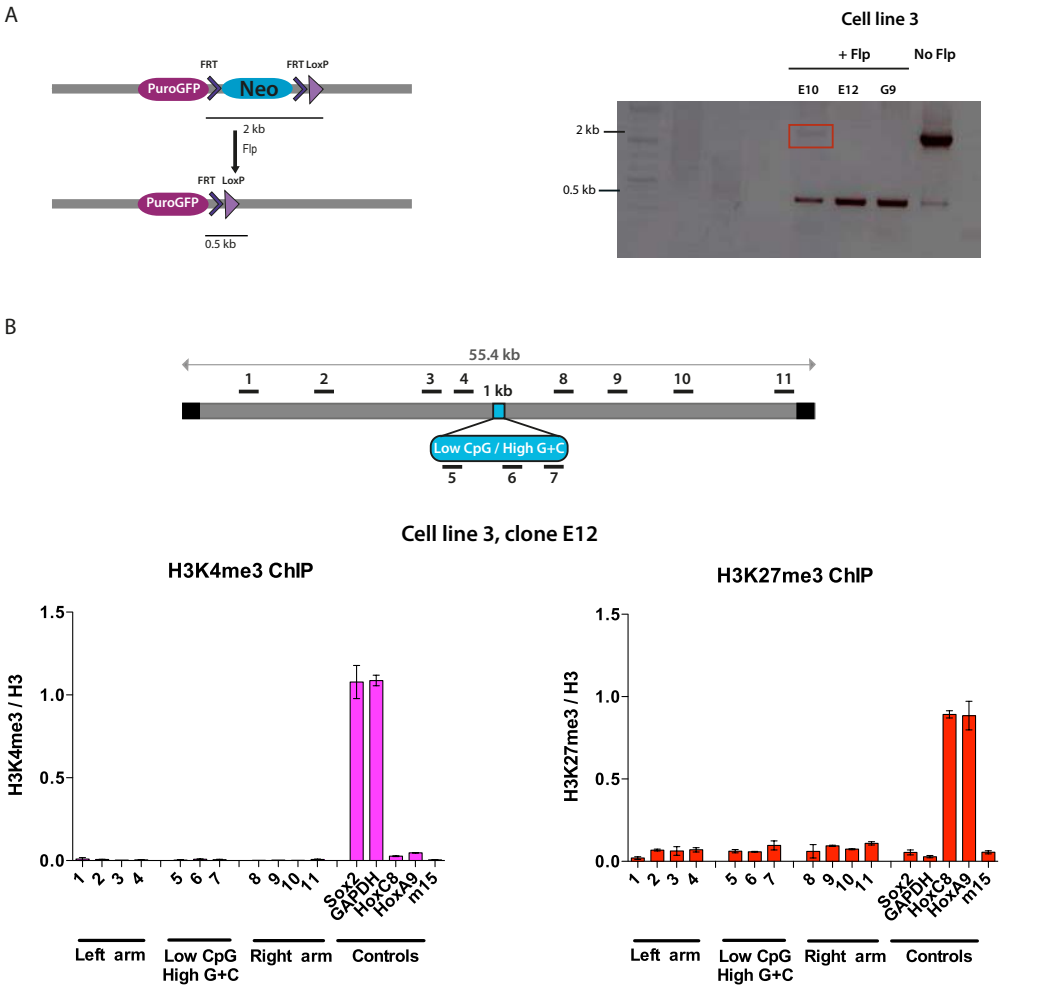


Figure 5.2.1-4 Presence of the selection cassettes disturbs analysis of Low CpG / High G+C influence on chromatin establishment
A: PCR screen to check for absence of selection cassette in different clones of cell line 3 after Flp transfection. Red box shows band that indicates incomplete excision of neo cassette. B: H3K4me3 and H3K27me3 ChIP of cell line 3 clone E12. Error bars indicate standard deviation of PCR replicates.

After showing that a high G+C content is not sufficient to create an H3K4me3 mark over the inserted region we wanted to know if it is adequate to recruit Polycomb proteins and establish the H3K27me3 mark. Figure 5.2.1-5 depicts the H3K27me3 and Suz12 ChIPs in all three cell lines. It is apparent that the introduced Lo/Hi sequence does not recruit Polycomb proteins, seen by the absence of Suz12 and H3K27me3 with levels in the region of those found in the negative control region m15. The bivalent control gene *HoxC8* shows as expected a good signal of both, H3K27me3 and Suz12 indicating the reliability of the ChIP experiment. This finding demonstrates that CpGs are not only essential for the recruitment of a H3K4 methyltransferase that mediates the establishment of H3K4me3 but also for the recruitment of Suz12. Interestingly, cell line 3, which was the only one showing an H3K4me3 peak, did not display an H3K27me3 signal as expected if its is due to the presence of the active neomycin gene as shown above (Figure 5.2.1-4).

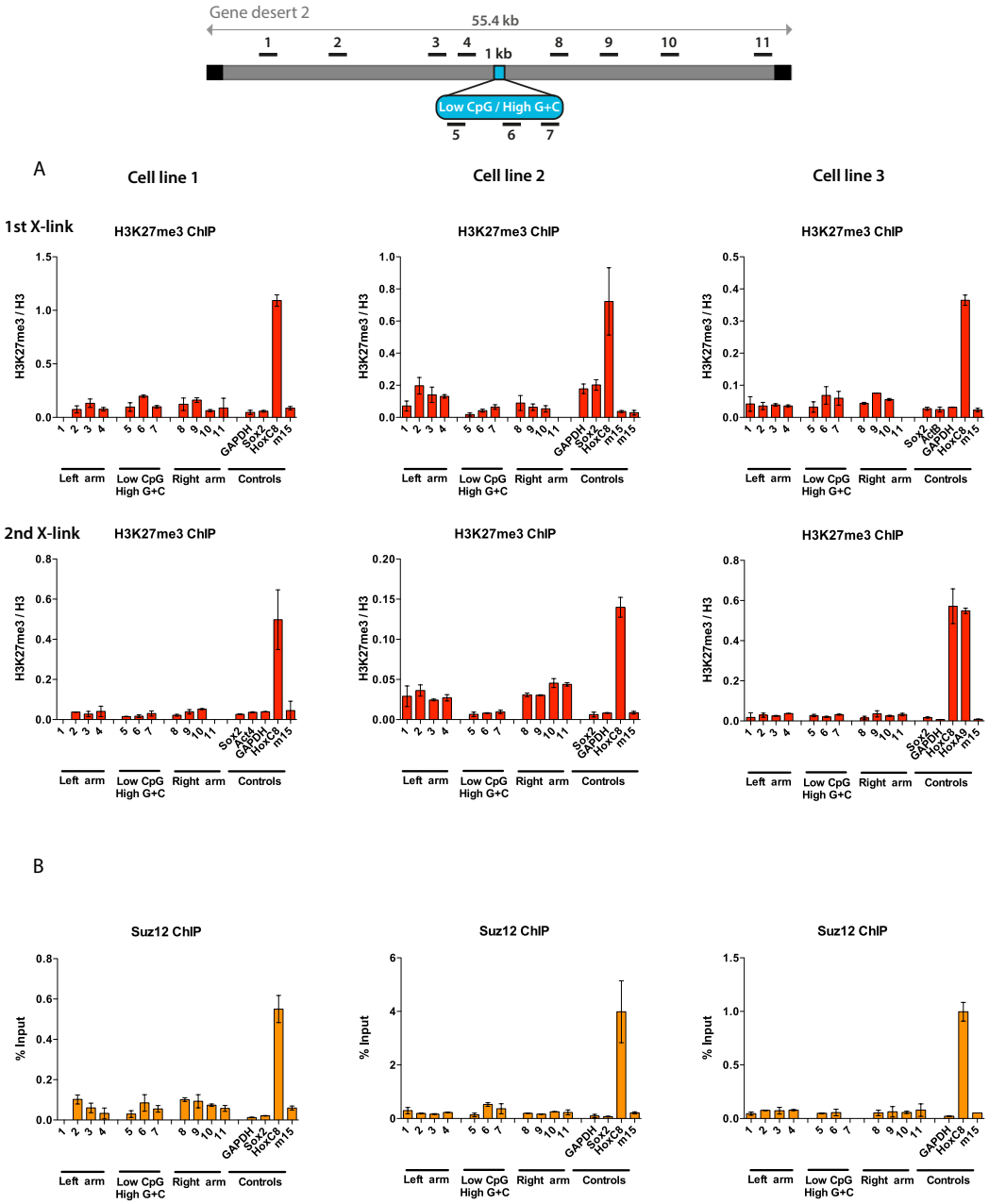


Figure 5.2.1-5 A high G+C content is not succificient to recruit Polycomb to Low CpG / High G+C construct in gene desert 2

A: H3K427me3 ChIP, 2 independent X-links for 3 cell lines. Y-axis: % of Input of H3K4me3 over % of Input of pan-H3 antibody and B: Suz12 ChIP, 1 X-link for 3 cell lines. Y-axis: % input. Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars indicate position of primers used for Q-PCR (length of amplicons not drawn to scale). Controls: TSS of active genes Sox2 and GAPDH, of bivalent gene HoxC8; m15= negative control region on mouse chromosome 15, indicates background levels. Error bars indicate standard deviation of PCR replicates.

5.2.2. High CpG frequency is not enough to protect a sequence from DNA methylation in mouse embryonic stem cells

Having shown that CpGs are important for the establishment of a bivalent domain we wanted to know if an overall high G+C content plays a role in attracting histone-modifying proteins as well. Therefore a sequence was designed that contains a high number of CpGs, similar to that conventionally found in CGIs but with a G+C content of 40,3%, which is comparable to that of the bulk genome, and much lower than the ~65% found in CGIs. Figure 5.2.1-1 shows the base composition of this High CpG / Low GC (also referred to as Hi/Lo) sequence in comparison to other sequences used in this study. The exact sequence can be found in the appendix. As before three independent ES cell lines that have integrated the Hi/Lo sequence in gene desert 2 were established. Figure 5.2.2-1 displays the number of integrations of the different BAC clones and indicated the 3 cell lines that were taken further in order to excise the selection cassette.

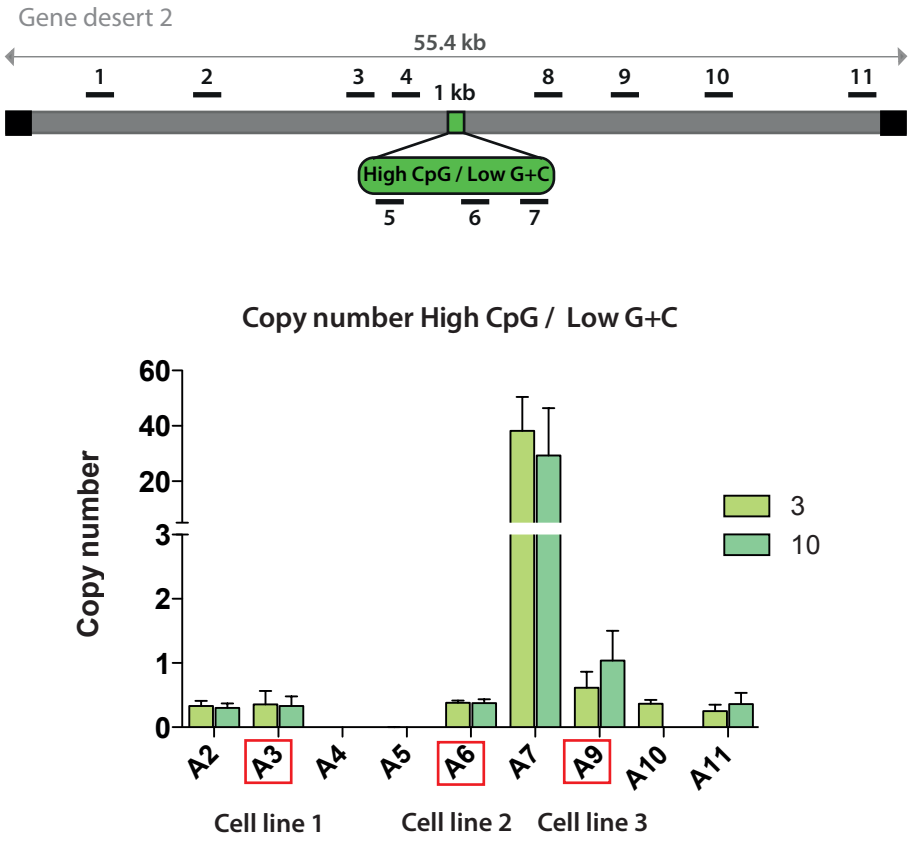


Figure 5.2.2-1 Copy number analysis of High CpG / Low G+C sequence in gene desert
Mouse ES cell clones with integrated High CpG / Low G+C construct were analysed by Q-PCR for copy number integration. 3 and 10 respectively refers to the primer pair used in the PCR. Values were normalized to that of Sox2 = 2 copies per cell. Red boxes indicate chosen cell lines. Error bars indicate standard deviation of PCR triplicates. Scheme above indicates construct inserted in gene desert 2 and position of primers (not drawn to scale).

Once the excision of the selection cassette was confirmed by Southern blot the clones were expanded and a ChIP with anti H3K4me3 antibodies was performed (Figure 5.2.2-2). Unexpectedly, no signal of H3K4me3 was detected over the inserted Hi/Lo sequence, despite the presence of a high density of CpGs. This result was confirmed in two independent biological replicate experiments for all three cell lines. The control genes showed the expected pattern of strong enrichment over the active genes *GAPDH* and *Sox2*, slight enrichments over the bivalent gene *HoxC8* and no enrichment over the negative control region m15. This result indicates that a high density of CpGs is not sufficient to recruit a H3K4 methyltransferase and establish H3K4me3 and that the A+T rich environment is for some reasons detrimental to its recruitment.

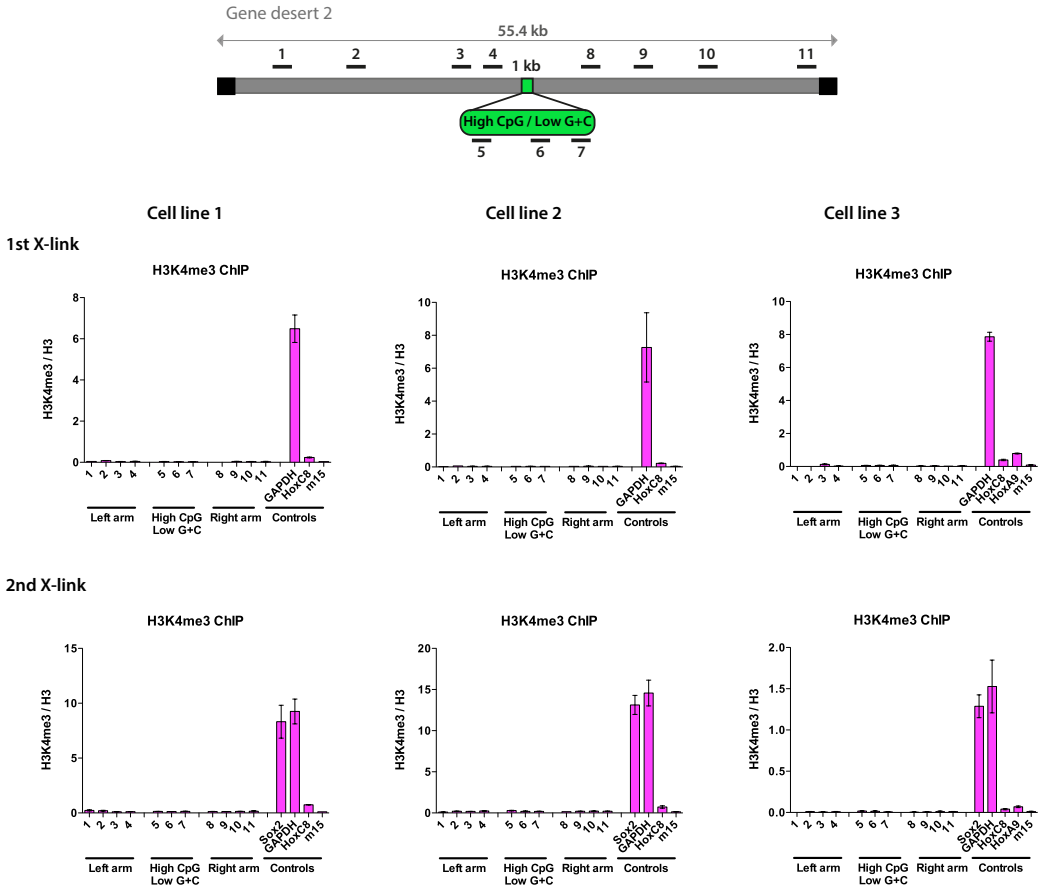


Figure 5.2.2-2 A high CpG density in an A+T rich background is insufficient to establish an H3K4me3 domain
H3K3me3 ChIP, 2 independent X-links for 3 cell lines. Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars denote position of primers used for Q-PCR (length of amplicons not drawn to scale). Y-axis: % of Input of H3K4me3 over % of Input of pan-H3 antibody. Controls: TSS of active genes Sox2 and GAPDH, of bivalent gene HoxC8; m15= negative control region on mouse chromosome 15, indicates background levels. Error bars indicate standard deviation of PCR replicates.

Next, I wanted to examine the effect of high CpG density but low overall G+C content on the recruitment of Polycomb as it is conceivable that a different mechanism is responsible for attracting the PRC2 complex. Figure 5.2.2-3 shows the result of the H3K27me3 and Suz12 ChIPs. Again, the signal over the inserted sequence shows no enrichment relative to flanking gene desert regions. Especially when looking at the Suz12 ChIP it seems that there is no enrichment over the inserted Hi/Lo sequence in comparison with the negative control region m15. In summary, we conclude that a high density of CpGs in an A+T rich environment is not sufficient to establish either of the characteristic histone marks of a bivalent domain.

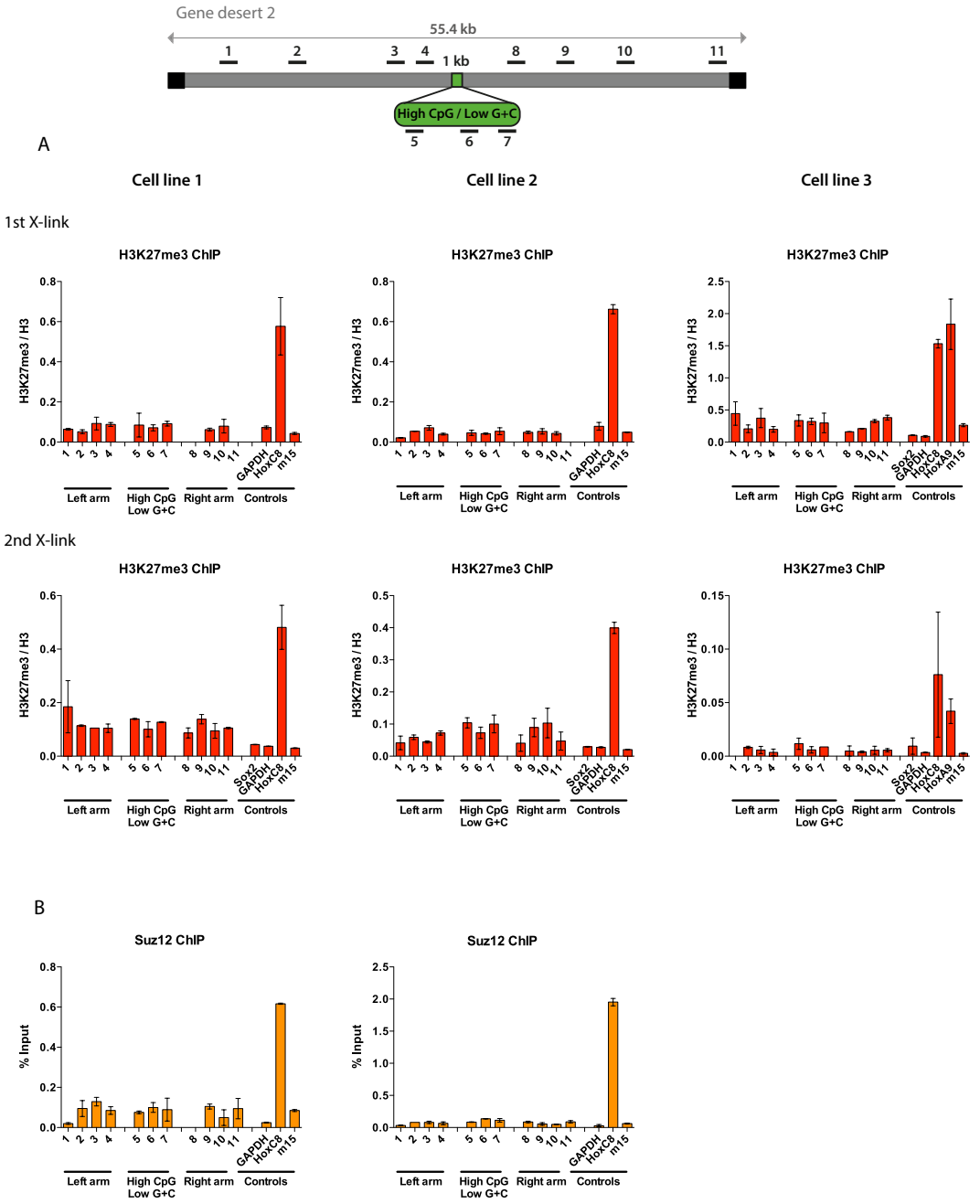


Figure 5.2.2-3 A high GpG content in an AT rich background is not sufficient to recruit Polycomb

A: H3K427me3 ChIP, 2 independent X-links for 3 cell lines. Y-axis: % of Input of H3K4me3 over % of Input of pan-H3 antibody. B: Suz12 ChIP, 1 X-link for 2cell lines. Y-axis: % input. Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars indicate position of primers used for Q-PCR (length of amplicons not drawn to scale). Controls: TSS of active genes Sox2 and GAPDH, of bivalent gene HoxC8; m15= negative control region on mouse chromosome 15, indicates background levels. Error bars indicate standard deviation of PCR replicates.

5.2.3. DNA methylation masks the effect of high CpG frequency in an A+T rich background on establishment of bivalent domain

We decided to investigate why the CpG rich but G+C poor sequence was not able to establish a bivalent domain. One hypothesis is that the CpGs in this combination become methylated. This theory was tested by bisulfite sequencing of an amplicon that spans around ¼ of the High CpG / Low G+C construct. Indeed Figure 5.2.3-1 shows that in all three cell lines this sequence becomes heavily methylated unlike the other artificial CGI like sequence that stayed almost completely unmethylated (Figure 4.2.2-1).

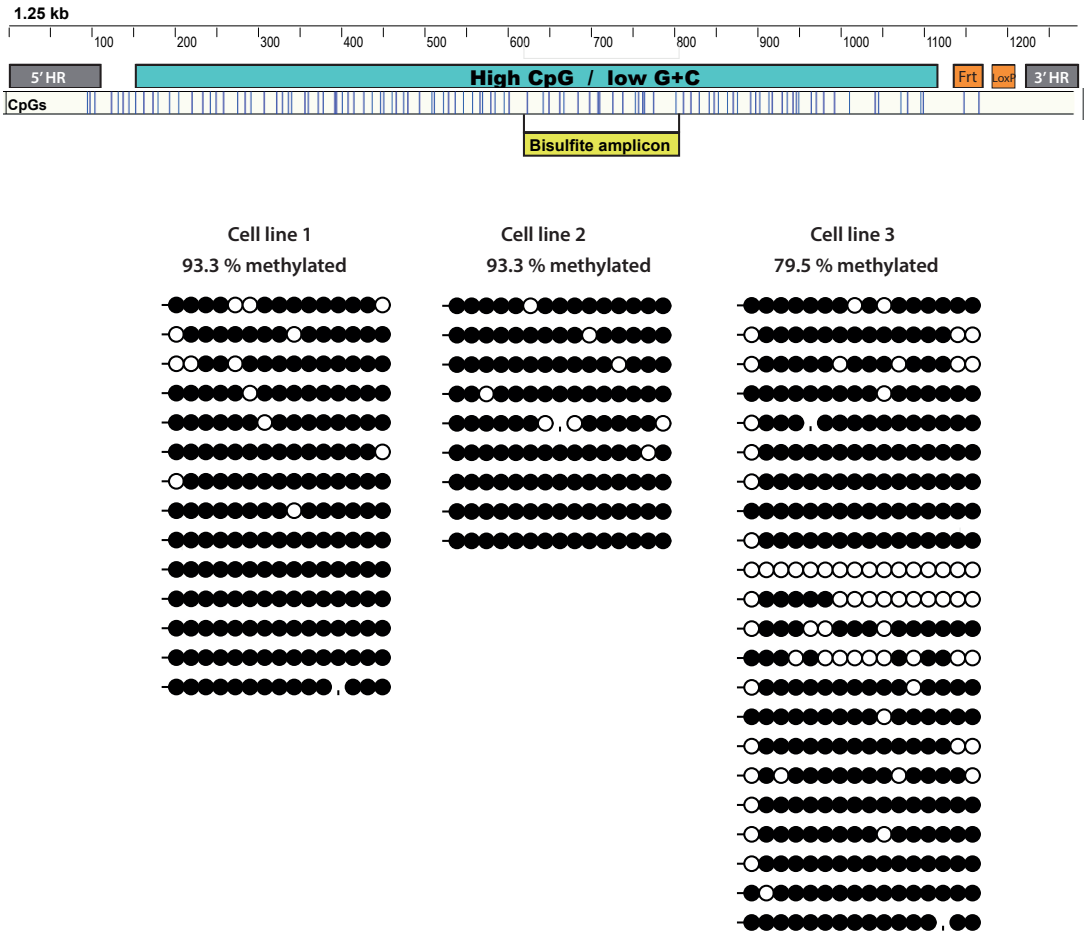


Figure 5.2.3-1 Construct with a high CpG density in an A+T rich environment becomes heavily methylated
Bisulfite sequencing of 3 cell lines containing the High CpG / Low G+C construct in gene desert 2. Scheme on top shows the inserted construct with the flanking gene desert arms and the remaining frt and LoxP sites. Scale is in bp. Blue vertical lines show position of CpGs. Sequenced amplicon is highlighted by yellow box. A methylated CpGs is depicted by a filled black circle, an unmethylated CpG by an empty white circle.

5.2.4. Masking effects of DNA methylation can be overcome using DNMT3a/3b double knock out mouse ES cells

In the previous section it was shown that a possible reason why a sequence with many CpGs but a low G+C content (High CpG / Low G+C) does not attract histone-modifying enzymes could be that they become heavily methylated. However, it is not clear if the High CpG / Low C+G sequence is intrinsically unable to form an H3K4m3 peak and therefore is getting methylated, as unmodified nucleosomes are good substrates for DNMTs. Or does this sequence become *de novo* methylated and therefore recruitment of histone modifying complexes is impeded, which results in the absence of H3K4me3? To test this question we analysed the Hi/Lo sequence in a system with reduced DNMT activity, which offers the possibility to investigate the effect of the Hi/Lo sequence on chromatin establishment without the masking effect of DNA methylation. This can be either achieved by knocking down the enzymes responsible for DNA methylation in the ES cell lines that have already integrated the sequence or treating them with a general demethylating agent. Alternatively, Dnmt knock out cells could be used for transformation with the BAC containing the Hi/Lo sequence. Treating cells with a demethylating agent such as 5-azacytidine has severe side effects as this agent has also DNA damaging effects. Moreover the demethylation achieved with this agent is usually incomplete (Yang *et al*, 2006; Palii *et al*, 2008). Knocking down the Dnmts by small interfering RNA (siRNA) would be another possibility. This would have the advantage that the existing cell lines with the integrated Hi/Lo sequence could be used for the siRNA knockdown. Therefore, no additional cell lines would need to be created, which is a time consuming process. However, at least the two *de novo* Dnmts 3a and 3b would need to be targeted and it is uncertain if residual activity would still methylate the inserted Hi/Lo sequence. Therefore, it was decided to use *Dnmt3a/3b* double knock out ES cells as it has been shown that they have almost no DNA methylation (Okano *et al*, 1999). Using *Dnmt1* and *Dnmt3a/3b* triple knock out ES cells would have been possible (Tsumura *et al*, 2006). However, these cells are already resistant to neomycin, puromycin, hygromycin and blasticidin, making further manipulation challenging. It has been shown that in the Dnmt3a/3b double knock out ES cells (DKO) the presence of the maintenance DNA methyltransferase Dnmt1 cannot ensure proper levels of DNA methylation in absence of the *de novo* Dnmts, that are usually highly expressed in ES cells (Okano *et al*, 1998b). Moreover, using the DKOs, which are resistant to neomycin, puromycin and hygromycin, allows the use of blasticidin for the introduction of CGI like sequences.

In order to establish if the DKOs are a suitable system to study the effect of the Hi/Lo sequence on chromatin establishment it was necessary to confirm that these cells display a low DNA methylation level. Figure 5.2.4-1 shows that while wt ES cells are heavily methylated (97.5 %), Intracisternal A-particle (IAP) transposable elements in DKOs are widely unmethylated (16 %) despite the fact that those elements are among the most highly methylated elements that are even resistant to the global demethylation wave during early embryonic development (Seisenberger *et al*, 2012).

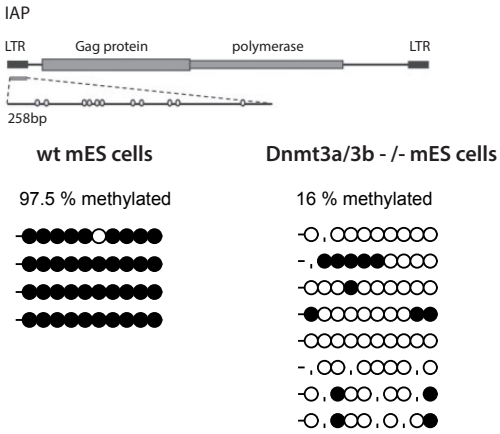


Figure 5.2.4-1 IAP elements are widely unmethylated in Dnmt3a/3b KO ES cells
Primers for bisulphite analysis are located in the 5'LTR of the IAP genome in order to amplify a 258-bp fragment containing up to 11 CpG dinucleotides spanning the IAP promoter (sequences obtained from (Lane *et al*, 2003)). A methylated CpGs is depicted by a filled black circle, an unmethylated CpG by an empty white circle.

Having shown that DNA methylation is strongly reduced at IAP elements we were confident that DKO ES cells could be used in order to answer the question of whether a High CpG / Low G+C sequence is able to attract histone modifying enzymes that establish H3K4me3 and/or H3K27me3. As can be seen in Figure 5.2.4-2 DKO ES cells seem grossly normal and comparable to the morphology of wt ES cells, although they show a little reduction in their growth rate and tend to grow more in patches.

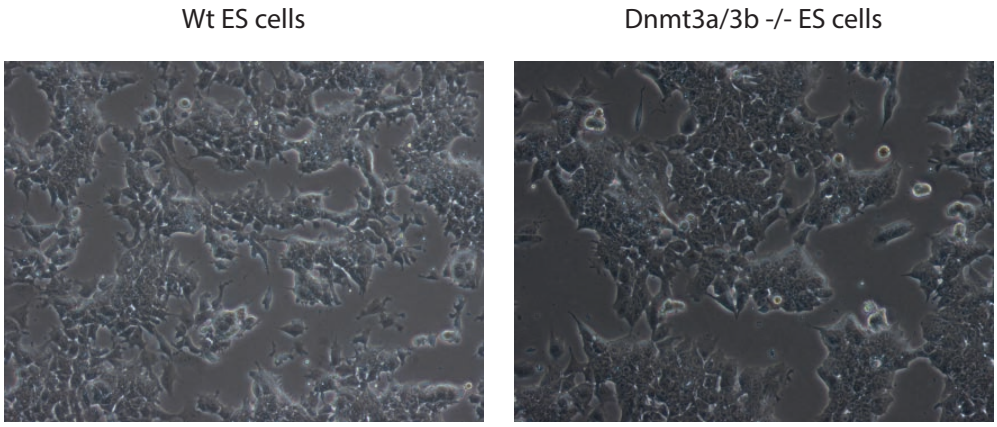


Figure 5.2.4-2 Dnmt3a/3b double knock out ES cells display similar morphology as wt ES cell
Representative pictures of wt versus Dnmt3a/3b KO ES cells cultured in ES medium containing 15 % FBS. 10 X magnification

As a first step we wanted to ask whether an artificial CGI-like sequence with high CpG frequency and a high G+C content can create a novel bivalent domain in a gene desert introduced into *Dnmt3a*/3b knock out cells. This control is designed to demonstrate that the observed creation of a bivalent domain of an artificial CGI-like sequence is not specific for wt ES cells but also occurs in the DKOs. Since the ArtCGI was already cloned into a vector containing a blasticidin resistance gene for the creation of the *Cfp1*-/- + ArtCGI and Cfp1-GFP + ArtCGI cell lines no new vector needed to be constructed. As described above DKOs were transfected with a gene desert containing BAC, where the ArtCGI and a blasticidin cassette were integrated via recombineering. Low Copy number integrations were selected and the clones expanded for transfection with DRE to excise the selection cassette. Clones were screened by PCR and Southern blot for the excision of the blasticidin cassette Positive clones were expanded and used for ChIP analysis with antibodies against H3K4me3 and H3K27me3.

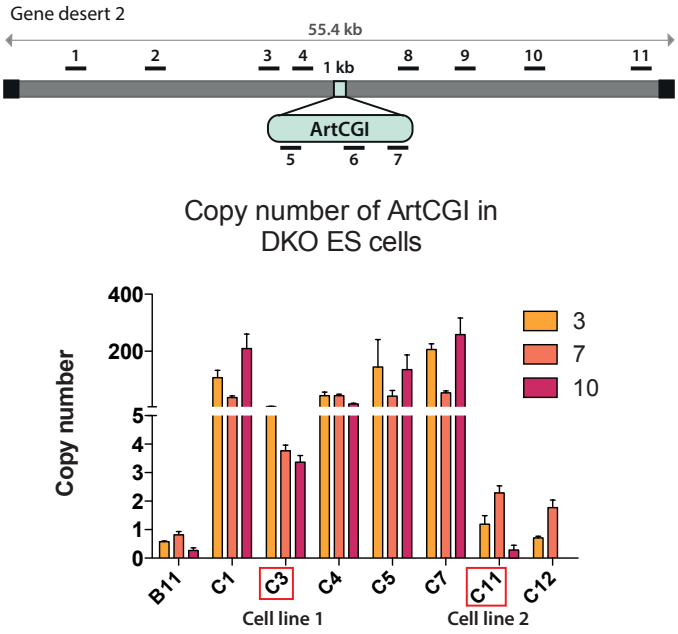


Figure 5.2.4-3 Copy number analysis of ArtCGI in gene desert in DKOs
Mouse ES cell clones with integrated ArtCGI construct were analysed by Q-PCR for copy number integration. 3, 7 and 10 respectively refers to the primer pair used in the PCR. Values were normalized to that of Sox2 = 2 copies per cell. Red boxes indicate chosen cell lines. Error bars indicate standard deviation of PCR triplicates. Scheme above indicates construct inserted in gene desert 2 and position of primers (not drawn to scale).

Figure 5.2.4-4 shows that the artificial CGI-like sequence integrated in a gene desert and randomly transfected into Dnmt3a/3b KO cells established a bivalent domain *de novo*. When the H3K4me3 signal over the ArtCGI is compared to the levels at the control genes, it becomes clear that it is a bit higher than over the bivalent genes *HoxC8*. Cell line 2 seems to have a stronger signal of H3K4me3 and H3K27me3 over the CGI in comparison to the bivalent control genes than cell line 1. But overall the chromatin established over the integrated CGI seems to resemble that of bivalent control genes.

In order to analyse the High CpG / Low G+C sequence it was necessary to perform a cassette exchange recombineering step that aimed to excise the neomycin gene and replace it with a blasticidin gene flanked by Rox sites, since the DKOs are already resistant to many antibiotics. As this cassette exchange proved to be straight forward for the ArtCGI construct (see 4.2.4) the same was expected for the Hi/Lo construct. However, despite many attempts no positive colonies that had undergone successful recombineering were obtained. It is possible that this additional recombineering step in an already 57kb big construct led to different rearrangements via non-homologous recombination. Despite the fact that care was taken to avoid gene desert regions with high amount of repetitive sequences for

recombineering this could not be excluded absolutely and might be one of the reasons why no positive clones were obtained.

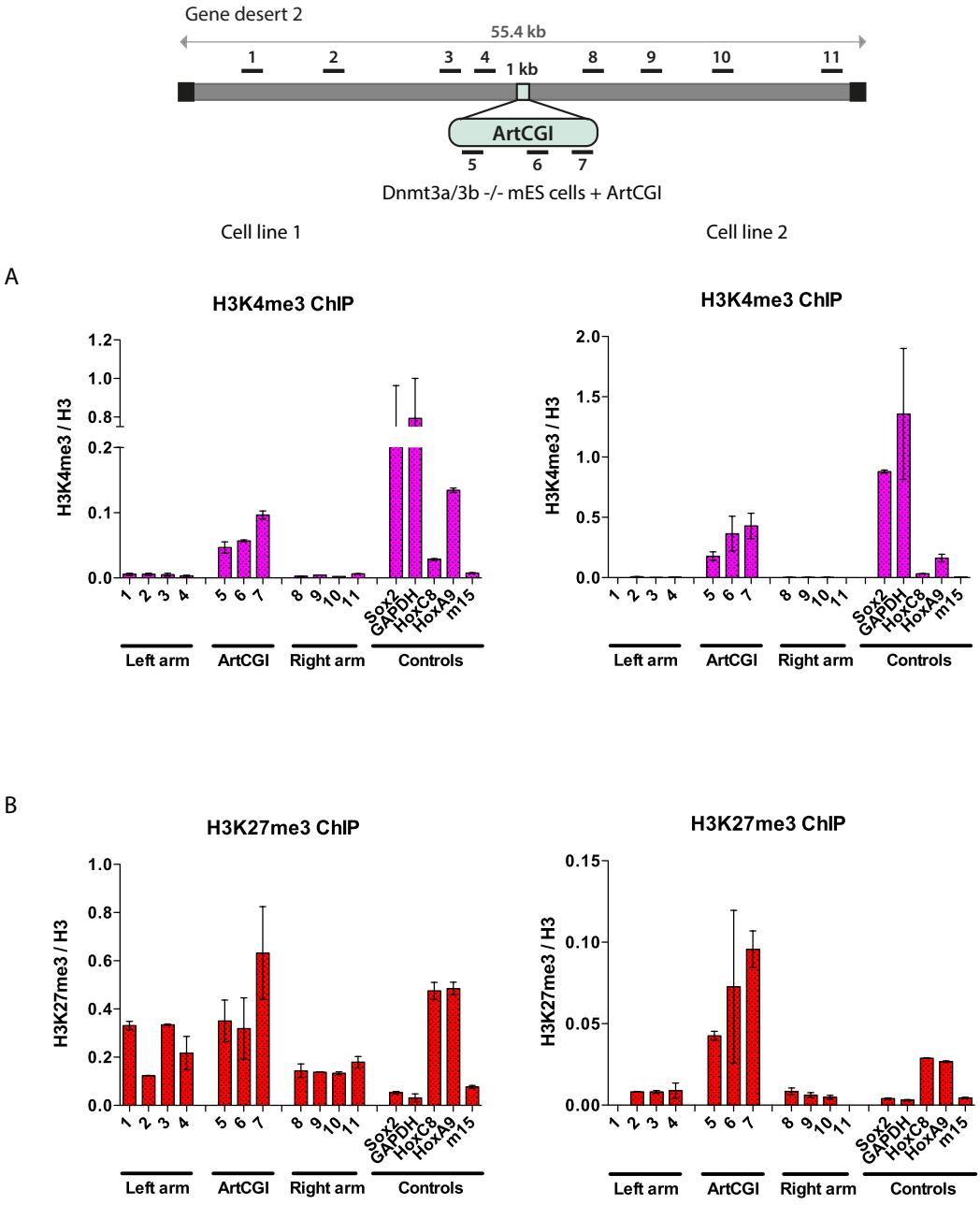


Figure 5.2.4-4 An artificial CGI-like sequence forms a bivalent domain in Dnmt3a/3b KO cells

Scheme on top indicates gene desert 2 (grey bar) and the integration site of CGI-like construct. Black boxes indicate bacterial backbone from BAC. Black bars indicate position of primers used for Q-PCR (length of amplicons not drawn to scale). A: H3K4me3 ChIP; B: H3K27me3 ChIP for 2 cell lines. Y-axis: % of Input of H3K4me3 or H3K27me3 over % of Input of pan-H3 antibody. Controls: TSS of active genes Sox2 and GAPDH and of bivalent genes HoxC8 and HoxA9; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates

As an alternative approach it was decided to create a new construct for BAC recombineering that contains the Hi/Lo sequence together with homology arms for the gene desert and the blasticidin resistance cassette. This construct can then be used to introduce into an unmodified gene desert 2 containing BAC. It was hoped that by using a fresh BAC the recombineering efficiency would rise. Unfortunately, despite optimizing different parameters like antibiotic concentration or method of amplifying constructs no positive clones were obtained. Ultimately a different strategy was adopted that involved co-transfection of DKO ES cells with a linearized gene desert 2 BAC and the Hi/Lo sequence only together with a plasmid containing the blasticidin selection cassette. It was assumed that those cells that would take up the plasmid carrying the resistance gene and therefore survive the drug selection would also take up the BAC with the Hi/Lo sequence. We hoped that by screening enough colonies these clones could be identified and used for analysing the effect of a High CpG low G+C sequence on establishment of chromatin without the influence of DNA methylation.

Although some colonies were obtained that were resistant to blasticidin they were much fewer in number than those obtained during other transformation experiments. The colonies were also smaller and less likely to survive picking. In the end around 70 colonies were available for screening and among them 1 colony was found to have integrated the BAC with the Hi/Lo sequence. As there was no selection cassette present within the BAC construct it was not necessary to excise the cassette and this one clone could be expanded and analysed directly. First a bisulfite sequencing experiment was performed to establish the methylation status of the High CpG / Low G+C sequence. Consistent with the fact that the two *de novo* Dnmts have been knocked out and even IAP elements proved to be largely unmethylated in these cells the Hi/Lo construct was found to be almost completely unmethylated, 0.05% methylation in DKOs versus around 80-90 % in wt ES cells (see Figure 5.2.4-5).

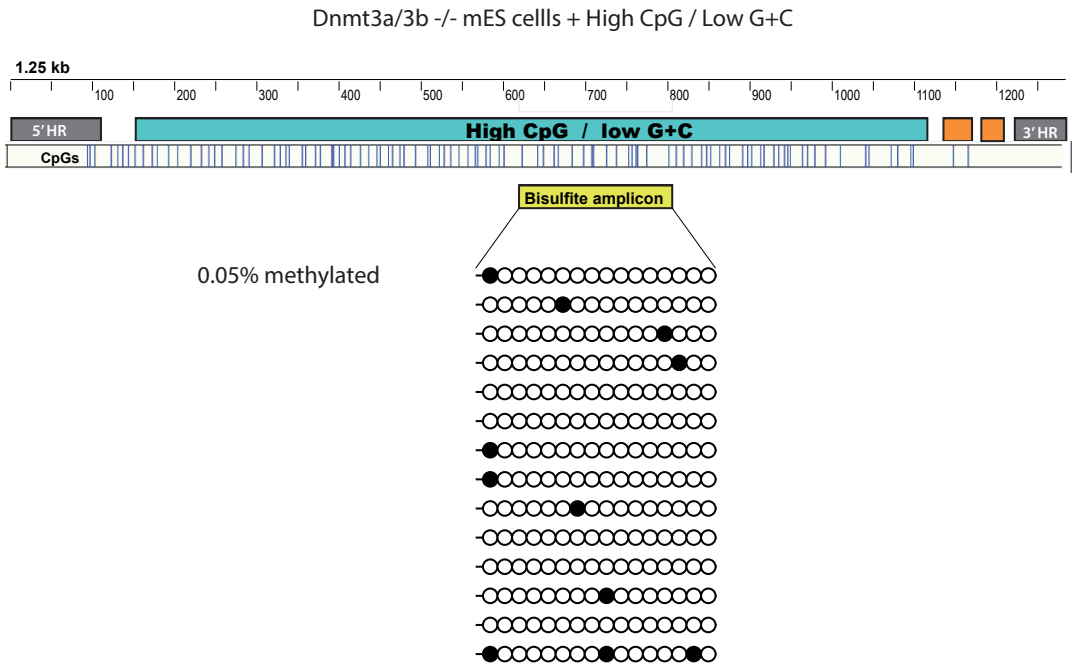


Figure 5.2.4-5 Construct with a high CpG density in an A+T rich environment stays unmethylated in Dnmt3a/3b KO cells
Bisulfite sequencing of the clone containing the High CpG / Low G+C construct in gene desert 2 in Dnmt3a/3b KO cells. Scheme on top shows the inserted construct with the flanking gene desert arms and the remaining frt and LoxP sites. Scale is in bp. Blue vertical lines show position of CpGs. Sequenced amplicon is highlighted by yellow box. A methylated CpGs is depicted by a filled black circle, an unmethylated CpG by an empty white circle.

Without the masking effects of DNA methylation it should now be possible to analyse the potential of the Hi/Lo sequence to establish a permissive or bivalent chromatin environment. Three independent X-link experiments were performed with the DKO + Hi/Lo cell line. Figure 5.2.4-6 shows the result of the ChIPs with H3K4me3 and H3K27me3 antibodies. The first ChIP experiment showed that the CpG rich sequence formed a novel peak of H3K4me3 over the integrated sequence that was not present at the adjacent gene desert to levels between the two bivalent control genes *HoxC8* and *HoxA9*. Also a H3K27me3 signal was clearly detected over the CpG rich sequence that spread into the adjacent gene desert just as seen for the artificial CGI (see for example Figure 4.2.1-5). However, when after this initial result the experiment was repeated in parallel for X-link 2 and 3 no clear peak of H3K4me3 could be detected at the High CpG / Low G+C construct. In contrast H3K27me3 was still present, similarly to X-link 1. It is not clear why there is this discrepancy between the H3K4me3 ChIP of X-link 1 and 2/3. It is conceivable that during the few (around 3) additional passages of the cells that occurred between X-link 1 and X-link 2/3 some residual DNA methylation was acquired that is detrimental to the recruitment of Cfp1 but that did not

influence the establishment of H3K27me3. Repetition of the ChIPs additionally to bisulfite analysis of the later passage clone would be necessary to unequivocally conclude if high CpG density is enough to recruit Cfp1 and establish a novel domain of H3K4me3. Moreover, it would be important to obtain additional cell lines to confirm the results.

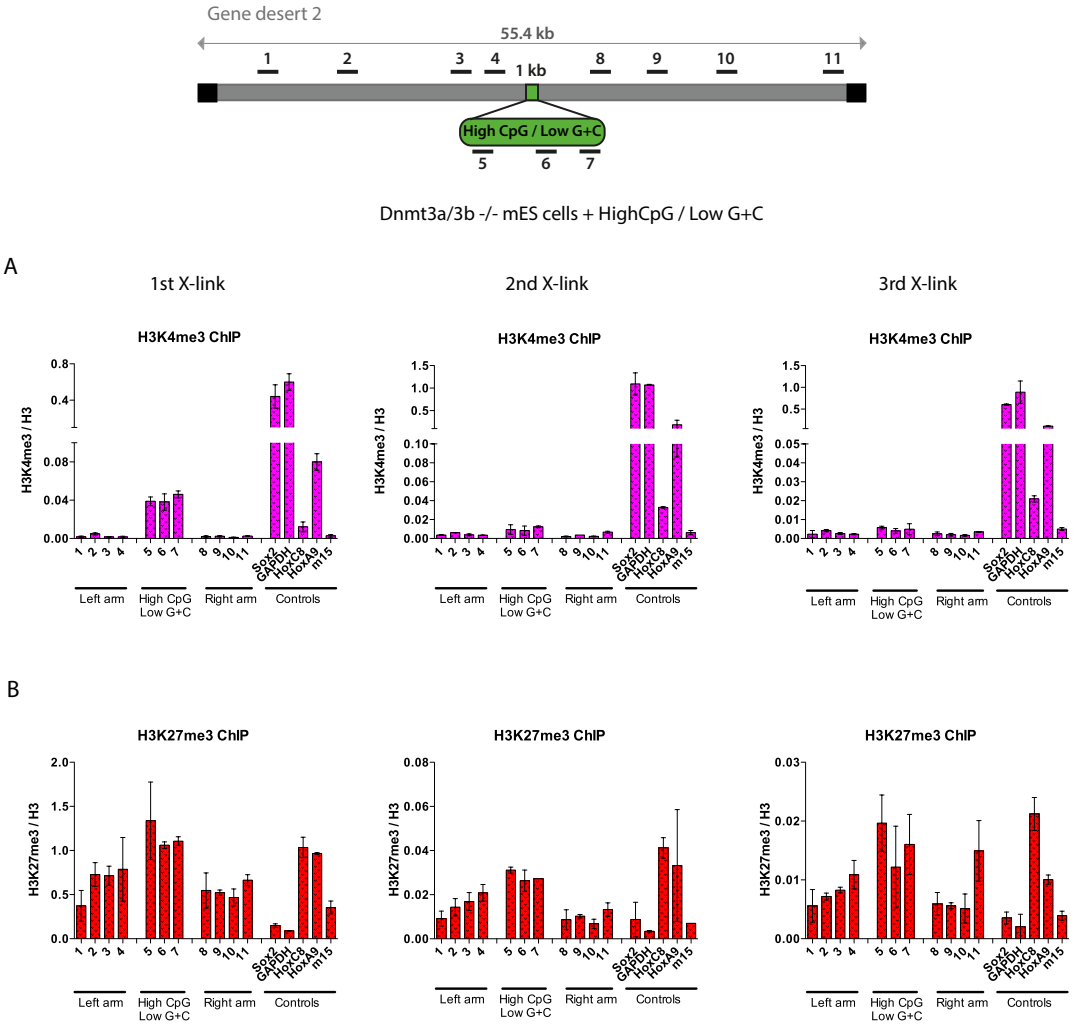


Figure 5.2.4-6 A bivalent domain is established over the High CpG / Low G+C sequence in Dnmt3a/3b KO ES cells
Scheme on top indicates gene desert 2 (grey bar) and the integration site of the High CpG / Low G+C construct. Black boxes indicate bacterial backbone from BAC. Black bars denote position of primers used for Q-PCR (length of amplicons not drawn to scale). A: H3K4me3 ChIP; B: H3K27me3 ChIP in 3 X-link experiments. Y-axis: % of Input of H3K4me3 or H3K27me3 over % of Input of pan-H3 antibody. Controls: TSS of active genes Sox2 and GAPDH and of bivalent genes HoxC8 and HoxA9; m15= negative control region on mouse chromosome 15 indicating background levels. Error bars indicate standard deviation of PCR replicates

5.3. Discussion

In this chapter it was shown that a sequence containing a low number of CpGs (like in the bulk genome) in a high GC rich environment is neither sufficient for H3K4me3 nor for H3K27me3 establishment. This result shows that a high density of CpGs is needed not only needed to attract Set1 but also for Polycomb recruitment. When looking at a sequence with the reverse base composition (high CpG density in an AT rich background) it was found that neither mark was created due to *de novo* DNA methylation of the inserted sequence. The finding that an overall high GC content is needed to protect a DNA sequence from becoming methylated is somewhat surprising as the sequence studied here contains a similar number of CpGs as found in most CGIs that stay constitutively unmethylated.

5.3.1. Why does a CpG rich sequence in a G+C poor (A+T rich) background become DNA methylated?

One possibility is that this sequence might resemble repetitive elements and might be targeted for *de novo* methylation as a genome defence mechanism. Indeed for example the consensus sequence for a full-length pericentric major satellite repeat (taken from (Bulut-Karslioglu *et al*, 2012)) contains a G+C content of 38%, which is very similar to that found on average in bulk genomic DNA outside CpG islands. However, the observed over expected CpG density of 0.88 in the major satellite repeat is much higher than that usual for the bulk genome (below 0.6) and resembles more the CpG density of a CGI, albeit a weak one. Both the artificial and the puroGFP tested in this study had an observed over expected CpG value of 1 and a G+C content of between 66 and 69%. In contrast, the High CpG / Low G+C sequence contained the roughly the same number of CpGs per 100bp but showed a G+C content of only 40%, which is similar to that found in the major satellite repeat. In vivo pericentromeric repeats are heavily methylated and failure to do so is linked to genomic instability (Ehrlich, 2003), offering an explanation for the observed de novo methylation of the Hi/Lo sequence.

How are DNMT3A and/or DNMT3B recruited to the inserted High CpG / Low G+C sequence? It has been reported that Suv39h-mediated H3K9 trimethylation can recruit Dnmt3b to major satellite repeats present in pericentric heterochromatin (Lehnertz *et al*, 2003). In addition G9a, a histone methyltransferase, has been implicated in recruiting DNMT3A and 3B independently of its methyltransferase activity, as a point mutation in the SET domain of G9a did impede heterochromatin formation but not de novo methylation (Epsztejn-Litman *et al*, 2008). It seems that the ankyrin domain of G9a physically interacts

with the catalytic domain of DNMT3A and 3B. SETDB1 another H3K9 methyltransferase was shown to interact with DNMT3A and 3B but not with DNMT1 (Li *et al*, 2006). The histone modification H3K9me3 is bound by proteins such as heterochromatin proteins 1 (HP1) thus forming a heterochromatic structure. It has been reported that HP1 directly interacts with DNMT1 resulting in increased DNA methylation (Smallwood *et al*, 2007). A recent study describes how a subset of all major H3K9 methyltransferases coexist in large multimeric complexes and how they are recruited to major satellite repeats (Fritsch *et al*, 2010). Given the resemblance of the High CpG / Low G+C sequence to major satellite repeat it is conceivable that the High CpG / Low G+C sequence attracts histone-modifying enzymes that either through mediation of H3K9 methylation or through direct interactions recruit DNMTs. Therefore it would be interesting to analyse if H3K9me3 is established over the Hi/Lo sequence or if the histone methylases Suv39 or G9a can be detected at that region.

5.3.2. Why are H3K4me3 and H3K27me3 establishment inhibited by DNA methylation?

In this chapter it was shown that a sequence with a high CpG frequency but low G+C content becomes heavily methylated (see Figure 5.2.3-1). The observed DNA methylation is the reason why neither H3K4me3 nor H3K27me3 were detected over the inserted sequence despite the presence of a high density of CpGs. This was determined by *Dnmt3a/3b* KO cells, which proved to be permissive for the creation of H3K4me3 and H3K27me3 (see Figure 5.2.4-6). So why is DNA methylation detrimental to the establishment of a bivalent domain? That DNA methylation and H3K3me3 establishment could be mutually exclusive is maybe not so surprising considering the fact that DNA methylation is regarded as a repressive mark, whereas H3K4me3 is considered to be an activating chromatin mark. In the case of H3K27me3 the situation is not as clear, as both systems, Polycomb and DNA methylation, are part of a cells silencing mechanism.

It has been suggested that DNA methylation inhibits H3K4me3 by using *Dnmt1* -/- fibroblasts, which showed only very low levels of DNA methylation. For many genes in these cells lack of DNA methylation was sufficient to partially establish novel H3K4me3 peaks at loci previously devoid of H3K4me3 (Lande-Diner *et al*, 2007). Another study suggested that H3K4 and DNA methylation are largely antagonistic during spermatogenesis (Brykczynska *et al*, 2010). This notion confirms findings in somatic cells. For example, Mohn and colleagues showed that those promoters that gained DNA methylation during differentiation from ES cells into neurons lost H3K4 methylation (Mohn *et al*, 2008). The

incompatibility between DNA methylation and H3K4me3 at CGIs has also been predicted based on studies in human primary cells (Weber *et al*, 2007). Also, DNA methylation and the presence of the H3K4me3 mark were found to be mutually exclusive in ES cells (Isagawa *et al*, 2011). The above-mentioned studies correlate DNA methylation with the lack of H3K4me3. However, the exact mechanism of how DNA methylation influences H3K4 methyltransferases is less clear. It is likely that methylated CpGs interfere with the recruitment of H3K4 methyltransferases, presumably through the abrogation of CxxC domain binding. As discussed in Chapter 1.2.5 and 1.3.2, the H3K4 methyltransferases MLL1 and MLL2 possess a CxxC domain that enables specific binding to unmethylated CpGs. Also SET1A and SET1B associate with a CxxC domain containing protein, Cfp1 (Long *et al*, 2013). Alternatively, methylated CpG moieties could attract histone lysine demethylases, which could contribute to the observed absence of H3K4me3 at methylated loci (Lande-Diner *et al*, 2007).

There have been numerous studies analysing the compatibility between Polycomb and DNA methylation. Several reports have shown antagonism or mutual exclusiveness between H3K27me3 and DNA methylation (Kondo *et al*, 2008; Lindroth *et al*, 2008; Bartke *et al*, 2010; Wu *et al*, 2010b). In one study prostate cancer cells were compared to normal prostate cells using ChIP-microarrays and around 5% of promoters were identified that gained H3K27me3 in cancer cells (Kondo *et al*, 2008). The authors found that genes enriched with H3K27me3 had no detectable DNA hypermethylation, and most genes showing DNA hypermethylation had no enrichment for H3K27me3 (Kondo *et al*, 2008).

However, some groups have reported that genes marked by PcG in embryonic stem cells are more susceptible to DNA methylation in cancer (Schlesinger *et al*, 2006; Ohm *et al*, 2007). It remains unclear whether this is an indirect effect due to for example low occupancy by transcription factors, or a direct consequence of PcG proteins recruiting DNA methyltransferases, as suggested by a previous report (Viré *et al*, 2006). Conversly, other studies were unable to confirm an effect of EZH2 on DNA methylation as knock down of EZH2 had no effect on silencing by DNA methylation (Kondo *et al*, 2008). During normal differentiation of ES cells to neural precursors it was shown that many gene sequences undergo *de novo* DNA methylation, and a large portion of these are initially marked by the Polycomb complex (Mohn *et al*, 2008; Meissner *et al*, 2008).

Another evidence for the antagonism between Polycomb and DNA methylation came from a study that analysed the imprinted *Rasgrf1* locus, which shows DNA methylation on the paternal allele and histone H3 lysine 27 trimethylation on the maternal allele at a differentially methylated domain (Lindroth *et al*, 2008). In this study it was shown that these two methylation marks are mutually antagonizing, whereby one blocks the placement of the other. Manipulations that cause aberrant changes in the levels of one of these marks had the opposite effect on the other mark (Lindroth *et al*, 2008).

A recent study used nucleosomes methylated on DNA and on histone H3 to identify crosstalk between these two distinct classes of modification by SILAC nucleosome affinity purification (SNAP) (Bartke *et al*, 2010). The authors found that the PRC2 complex, which recognizes H3K27 methylation, is negatively regulated by DNA methylation. They suggest that this may enable PRC2 to associate preferentially with a specific chromatin state that is not silenced completely and can respond to external stimuli, such as poised genes (Bartke *et al*, 2010). One report proposed that DNMT3A dependent non-promoter DNA methylation promotes expression of neurogenic genes in postnatal neural stem cells by functionally antagonizing Polycomb repression (Wu *et al*, 2010b). The author show that it is not the DNMT3A binding but the DNA methylation activity of this enzyme that antagonizes Polycomb binding (Wu *et al*, 2010b).

A very recent work presented sequential ChIP- bisulfite-sequencing (ChIP-BS-seq) as an approach to quantitatively assess DNA methylation patterns associated with chromatin modifications or chromatin-associated factors directly (Brinkman *et al*, 2012). Brinkman and colleagues found that H3K27me3 and DNA methylation are compatible throughout most of the genome, except for CpG islands, where these two marks are mutually exclusive (Brinkman *et al*, 2012). Upon loss of DNA methylation in *DNMT1* and *3A/3B* triple knock out (TKO) cells, accumulation of H3K27me3 in broad local enrichments and a decrease of sharp localized H3K27me3 peaks were observed suggesting that DNA methylation prevents H3K27me3 deposition locally and at a megabase scale (Brinkman *et al*, 2012).

From all these studies it seems that loci that show dense DNA methylation are incompatible with H3K27me3 whereas loci that are marked by H3K27me3 can preferentially gain DNA methylation in certain circumstances. How the observed mutual exclusiveness of DNA methylation and H3K27me3 within regions of high CpG density is achieved mechanistically is not yet clear. It is possible that, as with the recruitment of H3K4 methyltransferases,

methylated CpGs inhibit the recruitment of polycomb proteins. Direct evidence for this idea came from the fact that recognition of CpGs via the CxxC domain of KDM2B, a subunit of PRC1, is abrogated by DNA methylation (Farcas *et al*, 2012). As it has not been shown formally how PRC2 is recruited to its target genes the exact mechanism of how DNA methylation inhibits PRC2 is not known. However, the fact that PRC2 occupancy and unmethylated CGIs is highly correlated (Ku *et al*, 2008) makes it likely that unmethylated CpGs play a role in the recruitment of PRC2.

6. Discussion

6.1.1. A high CpG frequency in CGIs is required for establishment of bivalent domains

In this study it was shown that H3K4me3 and H3K27me3 are established *de novo* at a sequence in a gene desert region that shows a high CpG frequency and an overall high G+C content thereby creating a bivalent domain. This result indicates that the establishment of a bivalent domain is a default mechanism that can occur at CGI-like sequences in the absence of other cues, such as transcriptional activators. Additionally it was shown that the presence of a high CpG density is required for the recruitment of H3K4/K27 methyltransferases. A high G+C content alone is not sufficient for the establishment of a bivalent domain (See Figure 5.2.1-3 and Figure 5.2.1-5). This specific requirement for CpGs is in accordance with the view that many histone-modifying proteins, which are found at CGIs, possess a CxxC domain that enables the recognition of unmethylated CpGs.

Establishment of H3K4me3

It seems that a basal level of H3K4me3 is established in absence of transcription. The creation of the H3K4me3 is likely to be independent of the H3K4 methyltransferases SET1A/1B as cells deficient in *Cfp1*, a subunit of these complexes, are still able to write this chromatin mark (see Figure 4.2.5-1). This is consistent with the observation that SET1A/1B mediate the bulk of the H3K4 trimethylation in mammalian cells (Wu *et al*, 2008) and that depletion of *Cfp1* leads to loss of H3K4me3 only at active genes (Clouaire *et al*, 2012). However, Cfp1 is present at the integrated CGI like sequence (see Figure 4.2.4-2). Other H3K4 methyltransferases such as MLL1 or MLL2, which also possess a CxxC domain that could play a role in recruiting HMTs to unmethylated CpGs, might be responsible for the creation of the H3K4me3 peak observed at the PuroGFP insertion and the artificial CGI (Bach *et al*, 2009; Birke *et al*, 2002). Indeed, it was shown that MLL2 is the H3K4 methyltransferase in mammals that is responsible for trimethylating H3K4 at bivalent genes (Hu *et al*, 2013). This is consistent with the levels of H3K4me3 observed at the inserted CGI-like sequences that were lower than at the gene promoter of actively transcribed genes like *GAPDH* or *Sox2* and showed more similarity with those at the bivalent control genes *HoxC8* and *HoxA9*. For future experiments it will be interesting to investigate which H3K4 methyltransferase is responsible for H3K4me3 at artificial CGI-like sequences. MLL2 might be the best candidate as it has been shown that this enzyme is responsible for establishing H3K4m3 at

bivalent genes (Hu *et al*, 2013). Therefore, *MLL2* knock out ES cell could be used or *MLL2* could be knocked down using shRNAs in cells already deficient for *Cfp1*. Alternatively, the role of other H3K4 methyltransferases could be examined.

Establishment of H3K27me3

As well as H3K4me3, H3K27me3 was detected at the artificial CGI-like sequences to levels similar to those found at the promoter region of bivalent genes. This finding was strengthened by the presence of Suz12, a member of the PRC2 complex. As the core PRC2 complex does not contain any DNA-binding domains itself, the mechanisms of its recruitment to gene loci have remained elusive. Even though PRC2 localizes to unmethylated CGIs, no CxxC domain-containing proteins are known to interact with PRC2. JARID2, an interactor of PRC2, does not possess a CxxC domain but shows binding preference to CG rich DNA (Li *et al*, 2010). Equally, AEBP2, a zinc finger protein that co-localizes with PRC2 binds DNA with low specificity (Kim *et al*, 2009). Additionally, orthologues of Drosophila Polycomb-like (PCL), also interact with PRC2 and have been implicated in its recruitment (Margueron & Reinberg, 2011; Simon & Kingston, 2013). Alternatively to DNA binding, many subunits of the PRC2 interact with nucleosomes that promote recruitment to chromatin without relying on DNA sequence-specific binding. For example, the PRC2 subunits EED and RBAP46/48 bind to histones H3 and H4. EED also binds to H3K27me3. PCL proteins contain PHD and Tudor domains via which additional interactions with nucleosomes can be formed. It has been proposed that the sum of relatively weak unspecific interaction can in the end lead to PRC2 recruitment (Klose *et al*, 2013; Margueron & Reinberg, 2011; Voigt *et al*, 2013). Additionally, it has been suggested that rather than being actively recruited to specific loci, PRC2 is sampling potentially permissive regions and does bind unless excluded by antagonistic signals (Farcas *et al*, 2012; Klose *et al*, 2013; Voigt *et al*, 2013). This means that prevention of binding could be a way of regulating PRC2 recruitment. Furthermore, KDM2B-PRC1 complexes continually probe unmethylated CpG loci for their susceptibility to repression, and stable recruitment may further depend on pre-existing repressive determinants (Mendenhall *et al*, 2010; Farcas *et al*, 2012).

This view of PRC2 recruitment partially fits with the data presented in this thesis that show H3K27me3 establishment at an isolated CGI-like sequence. The presence of PRC2 might be owed to the fact that no excluding activities such as productive transcription prevent recruitment in a gene desert region. Similar observations have been made by showing that

introduction of ectopic CGIs is sufficient to recruit PRC2 as long as these sequences are devoid of activating signals (Lynch *et al*, 2011; Mendenhall *et al*, 2010). Lynch and colleagues showed that replacing the normally non-bivalent mouse α-globin locus with the CGI containing human α-globin locus created an ectopic bivalent domain (Yuan *et al*, 2012; Lynch *et al*, 2011). The authors further demonstrated that the presence of H3K27m3 diminishes with transcriptional activation.

However, as discussed above, it was shown in this work that CpGs are needed for the establishment of H3K27me3. Moreover, a high G+C content was not sufficient to recruit the PRC2 complex. This indicates that recruitment of different Polycomb subunits to generally G+C rich DNA, as was for example proposed for JARID2, is likely to be not sufficient for successful Polycomb recruitment but that some specific CpG binding activity is needed for the recruitment of PRC2.

In the future it might be interesting to investigate if and which PRC1 complex is present at the artificial CGI and to answer the question if KDM2B can be detected. It might be tempting to speculate that the PRC1 complex, for which a specific CpG binding activity has been shown in form of the CxxC domain of KDM2B, might be in certain cases responsible for the recruitment of PRC2. Maybe the PRC1 variant containing KDM2B is initially recruited to CGIs. This recruitment could lead to H2AK119ub and chromatin compaction, which might recruit PRC2, as it has been shown that PRC2 is stimulated by the presence of nucleosomes (Schmitges *et al*, 2011; Yuan *et al*, 2012). This model of PRC1 mediated PRC2 recruitment is however very speculative as there are not sufficient data to demonstrate its validity and it is likely that a range of mechanisms is contributing to PRC2 recruitment. Nonetheless, knockdown experiments of different PRC1 and PRC2 components could shed light on the order of recruitment at ectopically introduced CGIs.

Establishment of bivalent domains

In summary, these data indicate that unmethylated CpGs in CGIs mediate recruitment of both SET1/MLL and PRC2 activities, as both H3K4me3 and H3K27me3 are established at ectopically introduced CGIs. See Figure 6.1.1-1 for a potential model of how CGIs influence chromatin. Despite the fact that it has been argued that H3K4me3 and H3K27me3 cannot coexist because PRC2 is inhibited by active chromatin marks (Voigt *et al*, 2012; Schmitges *et al*, 2011) a recent study has proposed a model of how PRC2 recruitment to nucleosome containing H3K4me3 can be achieved (Voigt *et al*, 2012). The authors found that

nucleosomes with only one H3K4me3 mark could still be methylated by PRC2 whereas inhibition of PRC2 required the presence of H3K4me3 on both copies of H3 (Voigt *et al*, 2012; 2013). Thus, the resulting asymmetric conformation with H3K4me3 and H3K27me3 occupying opposite H3 tails allows the coexistence of active and repressive marks within single nucleosomes at bivalent loci (Voigt *et al*, 2012; Voigt *et al*, 2013). A way of how bivalent domains are established could be that CGI-like sequences attract MLL/SET1 complexes to create basal levels of H3K4me3. If additional activating cues trough binding of transcription factors, assembly of the transcription machinery and productive transcription occur, H3K4me3 is reinforced while PRC2 is excluded. In cases where no such additional activating forces occur, basal levels of H3K4me3 could be further diminished by recruiting an H3K4me3 demethylase that can compete with H3K4me3 deposition, leading to the removal of H3K4me3 from at least one copy of H3 in a nucleosome. This would allow the PRC2 complex to methylate H3K27 on the opposite H3 tail. PRC1 variants that are recruited to CGIs via the CxxC containing protein KDM2B could cooperate with PRC2 and reinforce each other as PRC1 compacts chromatin, which stimulates PRC2. Further, H3K27me3 demethylases could play a role in preventing excessive H3K27me3, thereby maintaining equilibrium of both marks at bivalent loci.

Fate of bivalent domains upon differentiation

For future experiments it would be interesting to further investigate the fate of bivalent domains upon differentiation. Therefore we want to design two constructs that resemble the CGI of the *Aprt* promoter, as the CGI of this gene is well studied (Macleod *et al*, 1994; Brandeis *et al*, 1994). One construct will have the same sequence as the *Aprt* CGI, whereas the other will contain the *Aprt* CGI with mutated SP1 binding sites. It is thought that the construct with the mutated SP1 sites will behave like the artificial CGI presented in this thesis and form a bivalent domain, as it contains a CpG rich and G+C rich sequence. Upon differentiation Polycomb is suggested to take over as activating cues are absent. In contrast, the Aprt CGI sequence with intact SP1 binding sites should loose Polycomb as it contains transcription factor binding sites and should recruit the transcriptional machinery. In this case differentiation is thought to result in the maintenance of the activating H3K4me3 and H3K27me3 would be expected to be absent. This would allow testing the hypothesis that CGIs are nucleating sites of bivalent domain in absence of transcription that can be resolved upon differentiation.

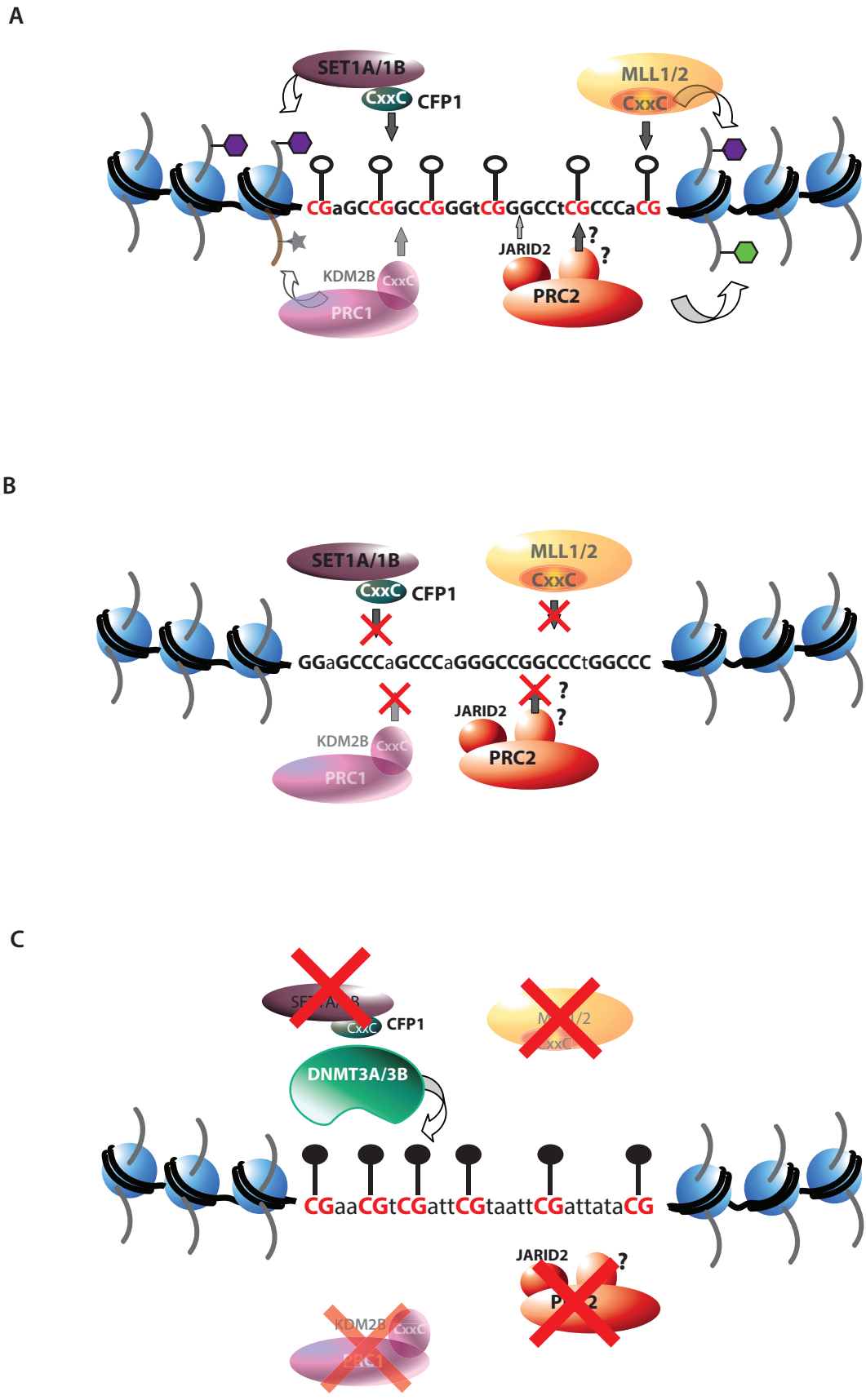

A
SET1A/1B
CxxC
CFP1
MLL1/2
CxxC
CGaGCCGGCCGGGtCGGGCCtCGCCCaCG
JARID2
?
?
KDM2B
CxxC
PRC1
PRC2
B
SET1A/1B
CxxC
CFP1
MLL1/2
CxxC
GGaGCCCaGCCCaGGGCCGGCCCtGGCCC
?
?
JARID2
KDM2B
CxxC
PRC1
PRC2
C
CFP1
DNMT3A/3B
CGaaCGtCGattCGtaattCGattataCG
JARID2
?
KDM2B
CxxC
H3K4me3
H3K27me3
H2AK119ub
Unmethylated CpG
Methylated CpG

Figure 6.1.1-1 Model of how CGIs influence chromatin
A: A CGI like sequence with high CpG density (Capital letters in red) and a high G+C content (Capital letters in black) in a gene desert creates a bivalent chromatin structure by default. This is characterized by the presence of H3K4me3, which might be established by MLL1/2 and/or SET1A/B, and by H3K27me3, which is established by the PRC2 complex. It is possible that the PRC1 complex might be recruited as well via the CxxC domain containing protein KDM2B. **B**: The presence of high CpG density is essential for H3K4 and H3K27 methyltransferases recruitment. A sequence that retains the same, high G+C content, but has very little CpGs (like the bulk genome) does neither attract H3K4me3 nor H3K27me3. **C:** A high G+C content is essential to protect a CGI from DNA methylation. A sequence that contains as many CpGs as an average CGI but has a C+G content like the bulk genome becomes heavily methylated, indicating that CpGs are not enough to protect against DNA methylation, which is detrimental to the establishment of both, H3K4me3 and H3K27me3. Complexes drawn in opaque have not been investigated in this study but might play a role.

6.1.1. A high G+C content in GGIs is required for maintaining their unmethylated state

In this thesis it was shown that a CGI-like sequence with a high CpG frequency and a high G+C content in a gene desert is sufficient to maintain an unmethylated state in mouse ES cells (see Figure 4.2.2-1). This result indicates that, at least in ES cells, transcriptional activity is not required for a CGI-like sequence to remain unmethylated, as no form of RNA polymerase II was detected at the inserted region. Moreover, it was demonstrated that a high G+C content is necessary to keep a sequence free of DNA methylation and that a high CpG content is not sufficient to prevent *de novo* DNA methylation (see Figure 5.2.3-1).

Interestingly, in the original study of Thomson and colleagues, where the promoterless CGI-like sequences eGFP and PuroGFP were introduced into the 3'end of the *Mecp2* and *Nanog* gene, respectively, eGFP became methylated whereas PuroGFP remained free of DNA methylation (Lienert *et al*, 2011; Thomson *et al*, 2010). Both CGI-like sequences display sequence properties that lie above the threshold that is usually applied for CGIs. It has been suggested that the locus of insertion might play a role. The eGFP sequence was integrated into the 3' UTR of Mecp2, a gene that is only lowly expressed in ES cells, whereas the PuroGFP was inserted into the 3'UTR of the highly expressed Nanog gene. However, it seems unlikely that the integration site is a factor influencing DNA methylation, as the same PuroGFP sequence remained unmethylated when integrated into a transcriptionally inert gene desert. When comparing the two CGI-like sequences differences become apparent; the eGFP sequence is shorter than the PuroGFP sequence (726 bp vs. 1287 bp). Moreover, both CpG frequency and C+G content are lower in the eGFP sequence than in the PuroGFP sequence (8 CpGs/100 bp and 61.6% G+C content vs. 11 CpGs/100 bp and 66.5% G+C

content). Together this indicates that eGFP is a "weaker" CGI and therefore might be more susceptible to DNA methylation.

A recent study by Lienert and colleagues, where around 50 different sequences were inserted into a transcriptionally inert locus, showed that these sequences autonomously recapitulated correct DNA methylation in absence of transcription (Lienert *et al*, 2011). However in this study they found that 7 out of 10 sequences from the *E.coli* genome that had an average length of 780 bp and varied in CpG density from 4.4 to 6.8 CpGs per 100 bp, became methylated. This CpG frequency is on the lower end of what normally constitutes a CGI but still well above the bulk genome. The 3 fragments that did not become methylated were among those with the higher frequency of CpGs, although some other fragments with the same CpG density did become methylated. On the first sight this observation might seem contradictory to the results obtained in this thesis that showed that a high CpG density and a high G+C content is enough to protect a sequence from becoming methylated. In order to investigate this issue further we analysed the exact sequence composition of the *E.coli* sequences used in the study by Lienert and colleagues. It became apparent that the G+C content of these sequences was considerably lower than that of average CGIs. The *E.coli* sequences displayed a G+C content between 40-55%, whereas usually CGIs have a G+C content of around 60%. The three fragments that did not become methylated showed the highest G+C content. This is in agreement with the finding presented here, which demonstrated that a high G+C content of 40% together with normal, high density of CpGs is not sufficient to protect against *de novo* DNA methylation. In order to further investigate the influence of G+C content on prevention of DNA methylation, a further construct will be created that displays the normal CpG density of a typical CGI but contains an intermediate G+C content of 51%.

In summary, CGIs are important features that are able to influence local chromatin to form a permissive structure that is per default characterized by absence of DNA methylation, presence of H3K4me3 and H3K27me3. DNA methylation was shown to be dominant over the formation of bivalent domains, presumably because recruitment of histone modifying enzymes is impaired by methylated Cs. Whereas CpGs are important for the recruitment of histone-modifying enzymes, a high G+C content is required to prevent DNA methylation (for a model see Figure 6.1.1-1). This shows how two common features of CGIs, CpG frequency and G+C content, act together in order to ensure an open chromatin structure that provides a platform for regulating transcription.

7. Appendix

PuroGFP sequence

GTCACCGAGCTGCAAGAACTCTTCCTCACGCGCGTCGGGCTCGACATCGGCAAG
GTGTGGGTCGCGGACGACGGCGCCGCGGTGGCGGTCTGGACCACGCCGGAGAG
CGTCGAAGCGGGGGCGGTGTTCGCCGAGATCGGCCCGCGCATGGCCGAGTTGAG
CGGTTCCCGGCTGGCCGCGCAGCAACAGATGGAAGGCCTCCTGGCGCCGCACCG
GCCCAAGGAGCCCGCGTGGTTCCTGGCCACCGTCGGCGTCTCGCCCGACCACCA
GGGCAAGGGTCTGGGCAGCGCCGTCGTGCTCCCCGGAGTGGAGGCGGCCGAGC
GCGCCGGGGTGCCCGCCTTCCTGGAGACCTCCGCGCCCCGCAACCTCCCCTTCTA
CGAGCGGCTCGGCTTCACCGTCACCGCCGACGTCGAGGTGCCCGAAGGACCGCG
CACCTGGTGCATGACCCGCAAGCCCGGTGCCTGACGCCCGCCCCACGACCCGCA
GCGCCCGACCGAAAGGAGCGCACGACCCCATGGCTCCGACCGAAGCCACCCGG
ATCCACCGGTCGCCACCATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGG
TGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGT
CCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCT
GCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCT
ACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCT
TCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGG
ACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTG
GTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTG
GGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGAC
AAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGA
CGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGG
CCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAA
AGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGC
CGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAA

Artificial CGI-like sequence

agcttCTAGCACAGGAGTTGACTCGGCAGGTGCCGTAGTGCGGCTTGGCAGCGGGA
CGTCGCTGGTCGGGCCTGACTGGGCGCTACGCCGGTTGTGGTCACCTGAACCGC
ATCTGGGCCGTCGCTCGCTTCGCGGGCTCTGCAGCCGGACTCCACCAGCGGGAC
CTCACACGCTCGGGTGAGCCGTCCTAGGCCGCTTGCGCCAACCCACGGGGTAGG
CCTGGCGAGACGCACGGGCAGTGCCGTTCCTGAGGTCCGCGGTGCTCCCTGCCC
AGACGCCTAAGCACGCTCCACCCGTGCCTCGGGTTCCGGGGATTTAGGCCGCAC

TGTGCGCCATTCCGCGGCTCGCAGCCACCGAAGTGCCGCGTTCCCTCCTACCCTA
GCCGCAGCAGCGTCCAGGGCCAAGAGGCTCGGCTACGGTCCGACTTTTGCACTG
GGCAGTCGCAGCGTCGTCCTCTCGCGACGCGGGTGCAGAGGTCGGCGGGTCGTA
ACCGACTGCCAACCAGGCGCAGGCTGGTACCCACCGGGTGGGCATAACGTCCAT
CGCCTGGGAGGTGGGACGCGGACTCGTCGGACACGTCGCCTCTGCGTGTGTCGC
GGGCCAGCAGCCAGCCGGCCTGGATCTAGACCCGGCGGCCATTGCAGCGAAGG
GAGGTTCGCCACGGCGTAGCCCAGTCGCGCGTGTACCCAGCGCATCCGCCGGAG
CCCATCCCCGCTCACGGTCGAAGCCCACCAGGGTGTCTCCACCGCGGCCTTTGCC
CACCTCGCTGGTCCACAGCCGAGCGGTCCGCTCTGACGCAGCGTCCTGCCAGCC
CTCGACGGCCGACCCCTGACTAGTTGGACCGCCTCCAGCGGCTGCGGGACACAG
CTCCGGCGGTGGGGACGCACCGAATGGGACACCGGCACTCCTAGGGCGTACAGC
TGCTCGCAACGCAACCGAAGGCAGCCTCCTACTGGCGGGCCGATCCGGCCTACA
TGGCCCTGGCCCATCATCCCGCGCAAGGTCGCACAATCGACGTGCACGACGGCA
GCACGCCAGGTGACGTTCCCGGAGCA

Low CpG / High G+C

GCTAAGCTTGGCCCCTGGGGGAAGCCCCACATGCCTGGGACCCCAAGCTGGACT
TCCGTCATGCAAGAGTACCAGGTACAGGGGACCCCTTAGTGTCTCCCTGGCAAG
TGCCCCCAGGGGACCAGGGTGGGCATCCCCCCACCTAAGTGGCCACCCCTCCCT
CCACCACTGTCCCCTGAAGGACATGTTGAGCCTGCCTCACCGGGGTGTGGTGGC
CACTGGGTCTCCAGGGCTTGCCAAGGGGTCAGAATACTTCCGGGGTGTCTACCC
ATCCCCACCCTAGGGGATGCTAAGCGGGGGTCCAGCCCTGCCCATTCCCCCAGG
GGTAGGAGGGGGTCCCTGGTGCCCCTCCCCAGAAAGGAGGCCAGGGGAGTGGG
GGGAGCCTAAACTAAACCCACTCAGCCCCTGGGCCCAGTTGGGAACAGATATGG
CTAGGCGGGGAGGCAGGGGGCTGATGAGGGGCCTGGTAAGGTCTCCCTCCCAA
GTGGGGGGGTGGATGGGCCCAGAATCCTATGAGGACACATATCTTTGACTGGGC
AGCTTCAGAGGGGTAGGGCCCTTGAGGGCTGGGTAGAGTCCCAAGGCCCCAGG
GGGGATGTGACCCCCCTATGTCCTCAGCCCCCCCAGACCACAGAGAGTTCAGGA
AGGAGGGTAGCCCCGCCTCTCCAGGGCAGGTGACCCAGGGCCCCCTGGTAACTG
GGGGGGGAGCACCCTCATTGAACCCCCCGAGCCCATGTCAGGTGGCAGCCACTC
CCAGCCAGAAGCCCTGAGGGCCCATCCCAGGTGGCCCCTAAGGGGGAGGGGGG
ATTCCCAGGAATATTCTCCCAGCTTCAGGGCCTCAGTGAGAATCATGAGGGGCC
CTGGCTCCCGCCATACCCCACCAGCATATGGCCTCTCCCCGGGTTCAGGGAGAA
ACCCAGTGGGGCAGCAAATTAAGCATCTCCCACTGACCAGAGCATTGGAGGTAG
GGGCTTCTGTAGGATGCCGGCCAAAGCTGCCAGCTGAACCCCTGAATTCTC

High CpG / Low G+C

CGTAAGCTTCTATAGCACACGGGCCAACGACAACGCTGGGCGATTTAACGTTTA
ATGTCGTATGAGTCTCGATAGCGAGGTTGGCACTCCGACCAAAAACCGAACATT
GAATCTAACGGAATACTGTCACGTTAATAACGTATAACGAAATATACGTATTTT
AACTATGAACGCATATAACGATTATCGCAGAACAAATTTTACGAATCAATAAAA
AACGTGATACGTAACTCGTTCGTTCAGAAGATTATGCGGACGAAAGAATGTACG
AACTCGTTGTATTTCTTGCGCGATATACGTAGTACGTGTTTCGTATGTAGTACCG
GAAAGTATCGAATCATTTCGATCGTACCTCACCGTTTCGACAACTACGATACGCT
AACAGTTGTTCGGCTATACAGCTTACGTCGCACAGTAGACGATTCCGACATGAC
GCAACTTATCGAAACAATTTCGATTTTAACGACGTACAAAATCGATTCGAATCA
CACTCGATATCGTACTATACAAATGTGAAATTCGGTCTCACTTTCTTGATCCGTT
TAGCGAATCTCTTTAACGCTTCGAGATTTAGTAGTTTTCGATGTAAATTTGACGA
AGTTTGTCGCGCATGAAAGAGTAAACGTCAATCTTCTCGATCTTATAACTATCGA
CCGAGCCGCGCCTTAGCTTCGCATATATGACAAATGACAAAACAAACGTGATTG
TCGCAGAATACGTTCTTTGACGTCATTTTAGACGAAATCGACTCGACTTCAATAC
GTTATACGATACGAATGTTGATCTGTTCGTGTACGCTTCGATCATTACTCGATCG
GGTAGTTGTGCGTTTTCGCATTGCGCACGACGAATTATCTGTAAACGTTTCCGGG
GGATGCGACCTCTCAGATCGTTATAAATGCTATTAACGTTATTATTCTTCCTTGG
ATAAGTCAAGTTCGAACGTGCTAAGAAGTGTGTTACTATATTACGAATGAACGA
GAGAGCATTAATACGACGATTCCCTAAATGAATTC

8. References

Ahmad K & Henikoff S (2002) The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Molecular Cell* **9:** 1191–1200

Akkers RC, Van Heeringen SJ, Jacobi UG, Janssen-Megens EM, Françoijs K-J, Stunnenberg HG & Veenstra GJC (2009) A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. *Developmental Cell* **17:** 425–434

Alder O, Lavial F, Helness A, Brookes E, Pinho S, Chandrashekran A, Arnaud P, Pombo A, O'Neill L & Azuara V (2010) Ring1B and Suv39h1 delineate distinct chromatin states at bivalent genes during early mouse lineage commitment. *Development* **137:** 2483–2492

Allen MD, Grummitt CG, Hilcenko C, Min SY, Tonkin LM, Johnson CM, Freund SM, Bycroft M & Warren AJ (2006) Solution structure of the nonmethyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase. *EMBO J* **25:** 4503–4512

Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U & Zoghbi HY (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* **23:** 185–188

Anastassiadis K, Fu J, Patsch C, Hu S, Weidlich S, Duerschke K, Buchholz F, Edenhofer F & Stewart AF (2009) Dre recombinase, like Cre, is a highly efficient site-specific recombinase in E. coli, mammalian cells and mice. *Disease Models & Mechanisms* **2:** 508–515

Ang Y-S, Tsai S-Y, Lee D-F, Monk J, Su J, Ratnakumar K, Ding J, Ge Y, Darr H, Chang B, Wang J, Rendl M, Bernstein E, Schaniel C & Lemischka IR (2011) Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. *Cell* **145:** 183–197

Antequera F & Bird A (1999) CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr Biol* **9:** R661–7

Ardehali MB, Mei A, Zobeck KL, Caron M, Lis JT & Kusch T (2011) Drosophila Set1 is the major histone H3 lysine 4 trimethyltransferase with role in transcription. *EMBO J* **30:** 2817–2828

Augui S, Nora EP & Heard E (2011) Regulation of X-chromosome inactivation by the X-inactivation centre. *Nat Rev Genet* **12:** 429–442

Ayton PM, Chen EH & Cleary ML (2004) Binding to nonmethylated CpG DNA is essential for target recognition, transactivation, and myeloid transformation by an MLL oncoprotein. *Mol Cell Biol* **24:** 10470–10478

Azuara V, Perry P, Sauer S, Spivakov M, Jørgensen HF, John RM, Gouti M, Casanova M, Warnes G, Merkenschlager M & Fisher AG (2006) Chromatin signatures of pluripotent cell lines. *Nat Cell Biol* **8:** 532–538

Bach C, Mueller D, Buhl S, Garcia-Cuellar MP & Slany RK (2009) Alterations of the CxxC

domain preclude oncogenic activation of mixed-lineage leukemia 2. *Oncogene* **28:** 815–823

Ball MP, Li JB, Gao Y, Lee J-H, LeProust EM, Park I-H, Xie B, Daley GQ & Church GM (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* **27:** 361–368

Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, Allshire RC & Kouzarides T (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410:** 120–124

Barski A, Cuddapah S, Cui K, Roh T, Schones D, Wang Z, Wei G, Chepelev I & Zhao K (2007) High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837

Bartke T, Vermeulen M, Xhemalce B, Robson SC, Mann M & Kouzarides T (2010) Nucleosome-Interacting Proteins Regulated by DNA and Histone Methylation. *Cell* **143:** 470–484

Bashtrykov P, Jankevicius G, Smarandache A, Jurkowska RZ, Ragozin S & Jeltsch A (2012) Specificity of Dnmt1 for methylation of hemimethylated CpG sites resides in its catalytic domain. *Chem Biol* **19:** 572–578

Baylin S & Herman J (2000) DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends Genet* **16:** 168–174

Baylin SB & Jones PA (2012) A decade of exploring the cancer epigenome-biological and translational implications. *Nature reviews cancer* **22:** 407–419

Bernstein BE, Meissner A & Lander ES (2007) The mammalian epigenome. *Cell* **128:** 669–681

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL & Lander ES (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125:** 315–326

Bestor TH (1992) Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain. *EMBO J* **11:** 2611–2617

Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev* **16:** 6–21

Bird A, Taggart M, Frommer M, Miller O & Macleod D (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40:** 91–99

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* **8:** 1499–1504

Bird AP (1995) Gene number, noise reduction and biological complexity. *Trends Genet* **11:** 94–100

Birke M, Schreiner S, García-Cuéllar M-P, Mahr K, Titgemeyer F & Slany RK (2002) The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and

discriminates against methylation. *Nucleic Acids Res* **30:** 958–965

Blackledge NP, Zhou JC, Tolstorukov MY, Farcas AM, Park PJ & Klose RJ (2010) CpG islands recruit a histone H3 lysine 36 demethylase. *Mol Cell* **38:** 179–190

Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, Ghavi-Helm Y, Wilczyński B, Riddell A & Furlong EEM (2012) Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat Genet* **44:** 148–156

Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, Levine SS, Wernig M, Tajonar A, Ray MK, Bell GW, Otte AP, Vidal M, Gifford DK, Young RA & Jaenisch R (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441:** 349–353

Boyes J & Bird A (1992) Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. *EMBO J* **11:** 327–333

Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, Temper V, Razin A & Cedar H (1994) Sp1 elements protect a CpG island from de novo methylation. *Nature* **371:** 435–438

Briggs SD, Bryk M, Strahl BD, Cheung WL, Davie JK, Dent SY, Winston F & Allis CD (2001) Histone H3 lysine 4 methylation is mediated by Set1 and required for cell growth and rDNA silencing in Saccharomyces cerevisiae. *Genes Dev* **15:** 3286–3295

Brinkman AB, Gu H, Bartels SJJ, Zhang Y, Matarese F, Simmer F, Marks H, Bock C, Gnirke A, Meissner A & Stunnenberg HG (2012) Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res* **22:** 1128–1138

Brykczynska U, Hisano M, Erkek S, Ramos L, Oakeley EJ, Roloff TC, Beisel C, Schübeler D, Stadler MB & Peters AHFM (2010) Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa. *Nat Struct Mol Biol* **17:** 679–687

Bulut-Karslioglu A, Perrera V, Scaranaro M, la Rosa-Velazquez de IA, van de Nobelen S, Shukeir N, Popow J, Gerle B, Opravil S, Pagani M, Meidhof S, Brabletz T, Manke T, Lachner M & Jenuwein T (2012) A transcription factor-based mechanism for mouse heterochromatin formation. *Nat Struct Mol Biol* **19:** 1023–1030

Butler JS, Lee J-H & Skalnik DG (2008) CFP1 interacts with DNMT1 independently of association with the Setd1 Histone H3K4 methyltransferase complexes. *DNA Cell Biol* **27:** 533–543

Cai L, Rothbart SB, Lu R, Xu B, Chen W-Y, Tripathy A, Rockowitz S, Zheng D, Patel DJ, Allis CD, Strahl BD, Song J & Wang GG (2013) An H3K36 methylation-engaging Tudor motif of polycomb-like proteins mediates PRC2 complex targeting. *Mol Cell* **49:** 571–582

Calo E & Wysocka J (2013) Modification of Enhancer Chromatin: What, How, and Why? *Molecular Cell* **49:** 825–837

Cao R & Zhang Y (2004) SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. *Molecular Cell* **15:** 57–67

Cao R, Tsukada Y-I & Zhang Y (2005) Role of Bmi-1 and Ring1A in H2A ubiquitylation and Hox gene silencing. *Molecular Cell* **20:** 845–854

Carlone DL & Skalnik DG (2001) CpG binding protein is crucial for early embryonic development. *Mol Cell Biol* **21:** 7601–7606

Carlone DL, Lee J-H, Young SRL, Dobrota E, Butler JS, Ruiz J & Skalnik DG (2005) Reduced genomic cytosine methylation and defective cellular differentiation in embryonic stem cells lacking CpG binding protein. *Mol Cell Biol* **25:** 4881–4891

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple C, Taylor M, Engstrom P, Frith M, Forrest A, Alkema W, Tan S, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, et al (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38:** 626–635

Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, Lee KK, Shia W-J, Anderson S, Yates J, Washburn MP & Workman JL (2005) Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123:** 581–592

Cedar H & Bergman Y (2009) Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet* **10:** 295–304

Chamberlain SJ, Yee D & Magnuson T (2008) Polycomb repressive complex 2 is dispensable for maintenance of embryonic stem cell pluripotency. *Stem Cells* **26:** 1496–1505

Chodavarapu RK, Feng S, Bernatavichute YV, Chen P-Y, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ, Casero D, Bernal M, Huijser P, Clark AT, Krämer U, Merchant SS, Zhang X, Jacobsen SE & Pellegrini M (2010) Relationship between nucleosome positioning and DNA methylation. *Nature* **466:** 388–392

Chuang LS, Ian HI, Koh TW, Ng HH, Xu G & Li BF (1997) Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. *Science* **277:** 1996–2000

Clouaire T, Las Heras de JI, Merusi C & Stancheva I (2010) Recruitment of MBD1 to target genes requires sequence-specific interaction of the MBD domain with methylated DNA. *Nucleic Acids Res* **38:** 4620–4634

Clouaire T, Webb S, Skene P, Illingworth R, Kerr A, Andrews R, Lee J-H, Skalnik D & Bird A (2012) Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes & Development* **26:** 1714–1728

Conerly ML, Teves SS, Diolaiti D, Ulrich M, Eisenman RN & Henikoff S (2010) Changes in H2A.Z occupancy and DNA methylation during B-cell lymphomagenesis. *Genome Res* **20:** 1383–1390

Cooper D, Taggart M & Bird A (1983) Unmethylated domains in vertebrate DNA. *Nucleic Acids Res* **11:** 647–658

Core LJ, Waterfall JJ & Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322:** 1845–1848

Cui K, Zang C, Roh T-Y, Schones DE, Childs RW, Peng W & Zhao K (2009) Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* **4:** 80–93

Daniels R, Kinis T, Serhal P & Monk M (1995) Expression of the myotonin protein kinase gene in preimplantation human embryos. *Hum Mol Genet* **4:** 389–393

Daniels R, Lowell S, Bolton V & Monk M (1997) Transcription of tissue-specific genes in human preimplantation embryos. *Hum. Reprod.* **12:** 2251–2256

Dawlaty MM, Breiling A, Le T, Raddatz G, Barrasa MI, Cheng AW, Gao Q, Powell BE, Li Z, Xu M, Faull KF, Lyko F & Jaenisch R (2013) Combined deficiency of Tet1 and Tet2 causes epigenetic abnormalities but is compatible with postnatal development. *Dev Cell* **24:** 310–323

Dawlaty MM, Ganz K, Powell BE, Hu Y-C, Markoulaki S, Cheng AW, Gao Q, Kim J, Choi S-W, Page DC & Jaenisch R (2011) Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development. *Cell Stem Cell* **9:** 166–175

De Smet C, Lurquin C, Lethe B, Martelange V & Boon T (1999) DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. *Mol Cell Biol* **19:** 7327–7335

Deaton AM & Bird A (2011) CpG islands and the regulation of transcription. *Genes & Development* **25:** 1010–1022

Deaton AM, Webb S, Kerr ARW, Illingworth RS, Guy J, Andrews R & Bird A (2011) Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome Research* **21:** 1074–1086

Delgado S, Gómez M, Bird A & Antequera F (1998) Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J* **17:** 2426–2435

Dhayalan A, Rajavelu A, Rathert P, Tamas R, Jurkowska RZ, Ragozin S & Jeltsch A (2010) The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *J Biol Chem* **285:** 26114–26120

Dinger M, Amaral P, Mercer T, Pang K, Bruce S, Gardiner B, Askarian-Amiri M, Ru K, Solda G, Simons C, Sunkin S, Crowe M, Grimmond S, Perkins A & Mattick J (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res* **18:** 1433–1445

Dou Y, Milne TA, Ruthenburg AJ, Lee S, Lee JW, Verdine GL, Allis CD & Roeder RG (2006) Regulation of MLL1 H3K4 methyltransferase activity by its core components. *Nat Struct Mol Biol* **13:** 713–719

Eberl HC, Spruijt CG, Kelstrup CD, Vermeulen M & Mann M (2013) A map of general and specialized chromatin readers in mouse tissues generated by label-free interaction proteomics. *Mol Cell* **49:** 368–378

Eckhardt F, Lewin J, Cortese R, Rakyan V, Attwood J, Burger M, Burton J, Cox T, Davies R, Down T, Haefliger C, Horton R, Howe K, Jackson D, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, et al (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* **38:** 1378–1385

Efroni S, Duttagupta R, Cheng J, Dehghani H, Hoeppner DJ, Dash C, Bazett-Jones DP, Le Grice S, McKay RDG, Buetow KH, Gingeras TR, Misteli T & Meshorer E (2008) Global transcription in pluripotent embryonic stem cells. *Cell Stem Cell* **2:** 437–447

Ehrlich M (2003) The ICF syndrome, a DNA methyltransferase 3B deficiency and immunodeficiency disease. *Clin Immunol* **109:** 17–28

Ehrlich M & Wang RY (1981) 5-Methylcytosine in eukaryotic DNA. *Science* **212:** 1350–1357

Ehrlich M, Gama-Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA & Gehrke C (1982) Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res* **10:** 2709–2721

Eissenberg JC & Shilatifard A (2010) Histone H3 lysine 4 (H3K4) methylation in development and differentiation. *Developmental Biology* **339:** 240–249

Endoh M, Endo TA, Endoh T, Isono K-I, Sharif J, Ohara O, Toyoda T, Ito T, Eskeland R, Bickmore WA, Vidal M, Bernstein BE & Koseki H (2012) Histone H2A mono-ubiquitination is a crucial step to mediate PRC1-dependent repression of developmental genes to maintain ES cell identity. *PLoS Genet* **8:** e1002774

Epsztejn-Litman S, Feldman N, Abu-Remaileh M, Shufaro Y, Gerson A, Ueda J, Deplus R, Fuks F, Shinkai Y, Cedar H & Bergman Y (2008) De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. *Nat Struct Mol Biol* **15:** 1176–1183

Erfurth FE, Popovic R, Grembecka J, Cierpicki T, Theisler C, Xia Z-B, Stuart T, Diaz MO, Bushweller JH & Zeleznik-Le NJ (2008) MLL protects CpG clusters from methylation within the Hoxa9 gene, maintaining transcript expression. *Proc Natl Acad Sci USA* **105:** 7517–7522

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M & Bernstein BE (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473:** 43–49

Eskeland R, Leeb M, Grimes GR, Kress C, Boyle S, Sproul D, Gilbert N, Fan Y, Skoultchi AI, Wutz A & Bickmore WA (2010) Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol Cell* **38:** 452–464

Estève P-O, Chin HG, Smallwood A, Feehery GR, Gangisetty O, Karpf AR, Carey MF & Pradhan S (2006) Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication. *Genes & Development* **20:** 3089–3103

Farcas AM, Blackledge NP, Sudbery I, Long HK, McGouran JF, Rose NR, Lee S, Sims D, Cerase A, Sheahan TW, Koseki H, Brockdorff N, Ponting CP, Kessler BM & Klose RJ (2012) KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. *elife* **1:** e00205

Feldman NN, Gerson AA, Fang JJ, Li EE, Zhang YY, Shinkai YY, Cedar HH & Bergman YY (2006) G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. *Nat Cell Biol* **8:** 188–194

Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I & Andrau J-C (2012) CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Research* **22:** 2399–2408

Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, Marques CJ, Andrews S & Reik W (2011) Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473:** 398–402

Filion GJ, Van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, De Castro IJ, Kerkhoven RM, Bussemaker HJ & Van Steensel B (2010) Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. *Cell* **143:** 212–224

Filion GJP, Zhenilo S, Salozhin S, Yamada D, Prokhortchouk E & Defossez P-A (2006) A family of human zinc finger proteins that bind methylated DNA and repress transcription. *Mol Cell Biol* **26:** 169–181

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Garcia-Giron C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kahari AK, Keenan S, Komorowska M, et al (2012) Ensembl 2013. *Nucleic Acids Res* **41:** D48–D55

Fossati A, Dolfini D, Donati G & Mantovani R (2011) NF-Y Recruits Ash2L to Impart H3K4 Trimethylation on CCAAT Promoters. *PLoS ONE* **6:** e17220–e17220

Fouse SD, Shen Y, Pellegrini M, Cole S, Meissner A, Van Neste L, Jaenisch R & Fan G (2008) Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. *Cell Stem Cell* **2:** 160–169

Francis NJ, Kingston RE & Woodcock CL (2004) Chromatin compaction by a polycomb group protein complex. *Science* **306:** 1574–1577

Frauer C, Rottach A, Meilinger D, Bultmann S, Fellinger K, Hasenöder S, Wang M, Qin W, Söding J, Spada F & Leonhardt H (2011) Different binding properties and function of CXXC zinc finger domains in Dnmt1 and Tet1. *PLoS ONE* **6:** e16627

Fritsch L, Robin P, Mathieu JRR, Souidi M, Hinaux H, Rougeulle C, Harel-Bellan A, Ameyar-Zazoua M & Ait-Si-Ali S (2010) A Subset of the Histone H3 Lysine 9 Methyltransferases Suv39h1, G9a, GLP, and SETDB1 Participate in a Multimeric Complex. *Molecular Cell* **37:** 46–56

Fujita N, Watanabe S, Ichimura T, Tsuruzoe S, Shinkai Y, Tachibana M, Chiba T & Nakao M (2003) Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. *J Biol Chem* **278:** 24132–24138

Gao ZZ, Zhang JJ, Bonasio RR, Strino FF, Sawai AA, Parisi FF, Kluger YY & Reinberg DD

(2012) PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. *Molecular Cell* **45:** 344–356

Gardiner-Garden M & Frommer M (1987) CpG Islands in vertebrate genomes. *Journal of Molecular Biology* **196:** 261–282

Gebhard C, Benner C, Ehrich M, Schwarzfischer L, Schilling E, Klug M, Dietmaier W, Thiede C, Holler E, Andreesen R & Rehli M (2010) General transcription factor binding at CpG islands in normal cells correlates with resistance to de novo DNA methylation in cancer cells. *Cancer Res* **70:** 1398–1407

Ginno PA, Lott PL, Christensen HC, Korf I & Chédin F (2012) R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Molecular Cell*

Glaser S, Schaft J, Lubitz S, Vintersten K, van der Hoeven F, Tufteland KR, Aasland R, Anastassiadis K, Ang S-L & Stewart AF (2006) Multiple epigenetic maintenance factors implicated by the loss of Mll2 in mouse development. *Development* **133:** 1423–1432

Gowher H, Liebert K, Hermann A, Xu G & Jeltsch A (2005) Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. *J Biol Chem* **280:** 13341–13348

Grau DJD, Chapman BAB, Garlick JDJ, Borowsky MM, Francis NJN & Kingston RER (2011) Compaction of chromatin by diverse Polycomb group proteins requires localized regions of high charge. *Genes & Development* **25:** 2210–2221

Gu T-P, Guo F, Yang H, Wu H-P, Xu G-F, Liu W, Xie Z-G, Shi L, He X, Jin S-G, Iqbal K, Shi YG, Deng Z, Szabó PE, Pfeifer GP, Li J & Xu G-L (2011) The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* **477:** 606–610

Guenther M, Levine S, Boyer L, Jaenisch R & Young R (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130:** 77–88

Guenther MG, Frampton GM, Soldner F, Hockemeyer D, Mitalipova M, Jaenisch R & Young RA (2010) Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell Stem Cell* **7:** 249–257

Guenther MG, Jenner RG, Chevalier B, Nakamura T, Croce CM, Canaani E & Young RA (2005) Global and Hox-specific roles for the MLL1 methyltransferase. *Proc Natl Acad Sci U S A* **102:** 8603–8608

Guo JU, Su Y, Zhong C, Ming G-L & Song H (2011a) Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain. *Cell* **145:** 423–434

Guo JU, Su Y, Zhong C, Ming G-L & Song H (2011b) Emerging roles of TET proteins and 5-hydroxymethylcytosines in active DNA demethylation and beyond. *Cell Cycle* **10:** 2662–2668

Gutierrez A (2004) Evolution of dnmt-2 and mbd-2-like genes in the free-living nematodes Pristionchus pacificus, Caenorhabditis elegans and Caenorhabditis briggsae. *Nucleic Acids Res* **32:** 6388–6396

Hackett JA, Sengupta R, Zylicz JJ, Murakami K, Lee C, Down TA & Surani MA (2013) Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. *Science* **339:** 448–452

Han H, Cortez CC, Yang X, Nichols PW, Jones PA & Liang G (2011) DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. *Hum Mol Genet* **20:** 4299–4310

Hansen KH, Bracken AP, Pasini D, Dietrich N, Gehani SS, Monrad A, Rappsilber J, Lerdrup M & Helin K (2008) A model for transmission of the H3K27me3 epigenetic mark. *Nat Cell Biol* **10:** 1291–1300

Hansen RSR, Stöger RR, Wijmenga CC, Stanek AMA, Canfield TKT, Luo PP, Matarazzo MRM, D'Esposito MM, Feil RR, Gimelli GG, Weemaes CMC, Laird CDC & Gartler SMS (2000) Escape from gene silencing in ICF syndrome: evidence for advanced replication time as a major determinant. *Hum Mol Genet* **9:** 2575–2587

Happel N & Doenecke D (2009) Histone H1 and its isoforms: Contribution to chromatin structure and function. *Gene* **431:** 1–12

Hathaway NA, Bell O, Hodges C, Miller EL, Neel DS & Crabtree GR (2012) Dynamics and memory of heterochromatin in living cells. *Cell* **149:** 1447–1460

He J, Shen L, Wan M, Taranova O, Wu H & Zhang Y (2013) Kdm2b maintains murine embryonic stem cell status by recruiting PRC1 complex to CpG islands of developmental genes. *Nat Cell Biol* **15:** 373–384

He Y-F, Li B-Z, Li Z, Liu P, Wang Y, Tang Q, Ding J, Jia Y, Chen Z, Li L, Sun Y, Li X, Dai Q, Song C-X, Zhang K, He C & Xu G-L (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333:** 1303–1307

Heard E, Rougeulle C, Arnaud D, Avner P, Allis CD & Spector DL (2001) Methylation of histone H3 at Lys-9 is an early mark on the X chromosome during X inactivation. *Cell* **107:** 727–738

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE & Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39:** 311–318

Hellman A & Chess A (2007) Gene body-specific methylation on the active X chromosome. *Science* **315:** 1141–1143

Hendrich B & Bird A (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol* **18:** 6538–6547

Hendrich B, Hardeland U, Ng HH, Jiricny J & Bird A (1999) The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* **401:** 301–304

Henikoff S & Shilatifard A (2011) Histone modification: cause or cog? *Trends Genet* **27:** 389–396

Hermann A, Gowher H & Jeltsch A (2004) Biochemistry and biology of mammalian DNA methyltransferases. *Cell Mol Life Sci* **61:** 2571–2587

Herz H-MH, Mohan MM, Garruss ASA, Liang KK, Takahashi Y-HY, Mickey KK, Voets OO, Verrijzer CPC & Shilatifard AA (2012) Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. *Genes & Development* **26:** 2604–2620

Ho L & Crabtree GR (2010) Chromatin remodelling during development. *Nature* **463:** 474–484

Hong S-H, Rampalli S, Lee JB, Mcnicol J, Collins T, Draper JS & Bhatia M (2011) Cell fate potential of human pluripotent stem cells is encoded by histone modifications. *Cell Stem Cell* **9:** 24–36

Hu D, Garruss AS, Gao X, Morgan MA, Cook M, Smith ER & Shilatifard A (2013) The Mll2 branch of the COMPASS family regulates bivalent promoters in mouse embryonic stem cells. *Nat Struct Mol Biol*

Hunkapiller J, Shen Y, Diaz A, Cagney G, McCleary D, Ramalho-Santos M, Krogan N, Ren B, Song JS & Reiter JF (2012) Polycomb-like 3 promotes polycomb repressive complex 2 binding to CpG islands and embryonic stem cell self-renewal. *PLoS Genet* **8:** e1002576

Iida T, Suetake I, Tajima S, Morioka H, Ohta S, Obuse C & Tsurimoto T (2002) PCNA clamp facilitates action of DNA cytosine methyltransferase 1 on hemimethylated DNA. *Genes Cells* **7:** 997–1007

Illingworth R, Kerr A, Desousa D, Jorgensen H, Ellis P, Stalker J, Jackson D, Clee C, Plumb R, Rogers J, Humphray S, Cox T, Langford C & Bird A (2008) A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol* **6:** e22

Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R & Bird AP (2010) Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet* **6:**

Ioshikhes IP, Albert I, Zanton SJ & Pugh BF (2006) Nucleosome positions predicted through comparative genomics. *Nat Genet* **38:** 1210–1215

Isagawa T, Nagae G, Shiraki N, Fujita T, Sato N, Ishikawa S, Kume S & Aburatani H (2011) DNA Methylation Profiling of Embryonic Stem Cell Differentiation into the Three Germ Layers. *PLoS ONE* **6:** e26052

Ito S, D'alessio AC, Taranova OV, Hong K, Sowers LC & Zhang Y (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466:** 1129–1133

Jia D, Jurkowska RZ, Zhang X, Jeltsch A & Cheng X (2007) Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature* **449:** 248–251

Jia J, Zheng X, Hu G, Cui K, Zhang J, Zhang A, Jiang H, Lu B, Yates J, Liu C, Zhao K & Zheng Y (2012) Regulation of pluripotency and self- renewal of ESCs through epigenetic-threshold modulation and mRNA pruning. *Cell* **151:** 576–589

Jiang H, Shukla A, Wang X, Chen W-Y, Bernstein BE & Roeder RG (2011) Role for Dpy-30 in ES Cell-Fate Specification by Regulation of H3K4 Methylation within Bivalent Domains. *Cell* **144:** 513–525

Jin C & Felsenfeld G (2007) Nucleosome stability mediated by histone variants H3.3 and H2A.Z. *Genes & Development* **21:** 1519–1529

Jin S-G, Kadam S & Pfeifer GP (2010) Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res* **38:** e125

Jones PA & Baylin SB (2007) The epigenomics of cancer. *Cell* **128:** 683–692

Jurkowska RZ, Jurkowski TP & Jeltsch A (2011) Structure and function of mammalian DNA methyltransferases. *Chembiochem* **12:** 206–222

Jørgensen HF, Ben-Porath I & Bird AP (2004) Mbd1 is recruited to both methylated and nonmethylated CpGs via distinct DNA binding domains. *Mol Cell Biol* **24:** 3387–3395

Kamakaka RT & Biggins S (2005) Histone variants: deviants? *Genes & Development* **19:** 295–310

Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, Li E & Sasaki H (2004) Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature* **429:** 900–903

Kanhere A, Viiri K, Araújo CC, Rasaiyaah J, Bouwman RD, Whyte WA, Pereira CF, Brookes E, Walker K, Bell GW, Pombo A, Fisher AG, Young RA & Jenner RG (2010) Short RNAs Are Transcribed from Repressed Polycomb Target Genes and Interact with Polycomb Repressive Complex-2. *Molecular Cell* **38:** 675–688

Karmodiya K, Krebs AR, Oulad-Abdelghani M, Kimura H & Tora L (2012) H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics* **13:** 424

Keshet I, Schlesinger Y, Farkash S, Rand E, Hecht M, Segal E, Pikarski E, Young RA, Niveleau A, Cedar H & Simon I (2006) Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat Genet* **38:** 149–153

Ketel CSC, Andersen EFE, Vargas MLM, Suh JJ, Strome SS & Simon JAJ (2005) Subunit contributions to histone methyltransferase activities of fly and worm polycomb group complexes. *Mol Cell Biol* **25:** 6857–6868

Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, Linder-Basso D, Plachetka A, Shanower G, Tolstorukov MY, Luquette LJ, Xi R, Jung YL, Park RW, Bishop EP, Canfield TK, et al (2011) Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471:** 480–485

Kim H, Kang K & Kim J (2009) AEBP2 as a potential targeting protein for Polycomb Repression Complex PRC2. *Nucleic Acids Res* **37:** 2940–2950

Kim T & Buratowski S (2009) Dimethylation of H3K4 by Set1 recruits the Set3 histone deacetylase complex to 5' transcribed regions. *Cell* **137:** 259–272

Klose RJ, Cooper S, Farcas AM, Blackledge NP & Brockdorff N (2013) Chromatin sampling-an emerging perspective on targeting polycomb repressor proteins. *PLoS Genet* **9:** e1003717

Koh KP, Yabuuchi A, Rao S, Huang Y, Cunniff K, Nardone J, Laiho A, Tahiliani M, Sommer CA, Mostoslavsky G, Lahesmaa R, Orkin SH, Rodig SJ, Daley GQ & Rao A (2011) Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell Stem Cell* **8:** 200–213

Kondo Y, Shen L, Cheng AS, Ahmed S, Boumber Y, Charo C, Yamochi T, Urano T, Furukawa K, Kwabi-Addo B, Gold DL, Sekido Y, Huang TH-M & Issa J-PJ (2008) Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation. *Nat Genet* **40:** 741–750

Kornblihtt AR (2006) Chromatin, transcript elongation and alternative splicing. *Nat Struct Mol Biol* **13:** 5–7

Kouzarides T (2007) Chromatin modifications and their function. *Cell* **128:** 693–705

Kriaucionis S & Heintz N (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324:** 929–930

Krogan NJ, Dover J, Khorrami S, Greenblatt JF, Schneider J, Johnston M & Shilatifard A (2002) COMPASS, a histone H3 (Lysine 4) methyltransferase required for telomeric silencing of gene expression. *J Biol Chem* **277:** 10753–10755

Krogan NJ, Dover J, Wood A, Schneider J, Heidt J, Boateng MA, Dean K, Ryan OW, Golshani A, Johnston M, Greenblatt JF & Shilatifard A (2003) The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation. *Molecular Cell* **11:** 721–729

Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS, Adli M, Kasif S, Ptaszek LM, Cowan CA, Lander ES, Koseki H & Bernstein BE (2008) Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet* **4:** e1000242

Kuzmichev A, Nishioka K, Erdjument-Bromage H, Tempst P & Reinberg D (2002) Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. *Genes & Development* **16:** 2893–2905

Lande-Diner L, Zhang J, Ben-Porath I, Amariglio N, Keshet I, Hecht M, Azuara V, Fisher AG, Rechavi G & Cedar H (2007) Role of DNA methylation in stable gene repression. *J Biol Chem* **282:** 12194–12200

Lander E, Linton L, Birren B, Nusbaum C, Zody M, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, et al (2001) Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921

Lane N, Dean W, Erhardt S, Hajkova P, Surani A, Walter J & Reik W (2003) Resistance of

IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *genesis* **35:** 88–93

Larsen F, Gundersen G & Prydz H (1992a) Choice of enzymes for mapping based on CpG islands in the human genome. *Genet Anal Tech Appl* **9:** 80–85

Larsen F, Gundersen G, Lopez R & Prydz H (1992b) CpG islands as gene markers in the human genome. *Genomics* **13:** 1095–1107

Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J & Wei C-L (2010) Dynamic changes in the human methylome during differentiation. *Genome Res* **20:** 320–331

Lee J-H & Skalnik DG (2005) CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J Biol Chem* **280:** 41725–41731

Lee J-H & Skalnik DG (2008) Wdr82 is a C-terminal domain-binding protein that recruits the Setd1A Histone H3-Lys4 methyltransferase complex to transcription start sites of transcribed human genes. *Mol Cell Biol* **28:** 609–618

Lee J-H, Tate CM, You J-S & Skalnik DG (2007a) Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex. *J Biol Chem* **282:** 13419–13428

Lee J-S, Shukla A, Schneider J, Swanson SK, Washburn MP, Florens L, Bhaumik SR & Shilatifard A (2007b) Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS. *Cell* **131:** 1084–1096

Lee J-S, Smith E & Shilatifard A (2010) The Language of Histone Crosstalk. *Cell* **142:** 682–685

Lee JH, Voo KS & Skalnik DG (2001) Identification and characterization of the DNA binding domain of CpG-binding protein. *J Biol Chem* **276:** 44669–44676

Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K-I, Koseki H, Fuchikami T, Abe K, Murray HL, Zucker JP, Yuan B, Bell GW, Herbolsheimer E, Hannett NM, Sun K, et al (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125:** 301–313

Leeb M, Pasini D, Novatchkova M, Jaritz M, Helin K & Wutz A (2010) Polycomb complexes act redundantly to repress genomic repeats and genes. *Genes Dev* **24:** 265–276

Lehnertz B, Ueda Y, Derijck AAHA, Braunschweig U, Perez-Burgos L, Kubicek S, Chen T, Li E, Jenuwein T & Peters AHFM (2003) Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr Biol* **13:** 1192–1200

Lei H, Oh SP, Okano M, Jüttermann R, Goss KA, Jaenisch R & Li E (1996) De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development* **122:** 3195–3205

Li B, Jackson J, Simon MD, Fleharty B, Gogol M, Seidel C, Workman JL & Shilatifard A (2009) Histone H3 lysine 36 dimethylation (H3K36me2) is sufficient to recruit the Rpd3s histone deacetylase complex and to repress spurious transcription. *J Biol Chem* **284:** 7970–7976

Li E, Bestor T & Jaenisch R (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69:** 915–926

Li G, Margueron R, Ku M, Chambon P, Bernstein BE & Reinberg D (2010) Jarid2 and PRC2, partners in regulating gene expression. *Genes & Development* **24:** 368–380

Li H, Rauch T, Chen Z-X, Szabó PE, Riggs AD & Pfeifer GP (2006) The histone methyltransferase SETDB1 and the DNA methyltransferase DNMT3A interact directly and localize to promoters silenced in cancer cells. *J Biol Chem* **281:** 19489–19500

Lienert F, Wirbelauer C, Som I, Dean A, Mohn F & Schübeler D (2011) Identification of genetic elements that autonomously determine DNA methylation states. *Nat Genet* **43:** 1091–1097

Lindroth AM, Park YJ, McLean CM, Dokshin GA, Persson JM, Herman H, Pasini D, Miró X, Donohoe ME, Lee JT, Helin K & Soloway PD (2008) Antagonism between DNA and H3K27 methylation at the imprinted Rasgrf1 locus. *PLoS Genet* **4:** e1000145

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B & Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462:** 315–322

Lo SM, Follmer NE, Lengsfeld BM, Madamba EV, Seong S, Grau DJ & Francis NJ (2012) A Bridging Model for Persistence of a Polycomb Group Protein Complex through DNA Replication In Vitro. *Molecular Cell* **46:** 784–796

Lock LF, Takagi N & Martin GR (1987) Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell* **48:** 39–46

Long HK, Blackledge NP & Klose RJ (2013) ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochem Soc Trans* **41:** 727–740

Lorincz MC, Dickerson DR, Schmitt M & Groudine M (2004) Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol* **11:** 1068–1075

Luger K, Mäder AW, Richmond RK, Sargent DF & Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389:** 251–260

Lyko F, Ramsahoye BH & Jaenisch R (2000) Development: DNA methylation in Drosophila melanogaster. *Nature* **408:** 538–540

Lynch MD, Smith AJH, De Gobbi M, Flenley M, Hughes JR, Vernimmen D, Ayyub H, Sharpe JA, Sloane-Stanley JA, Sutherland L, Meek S, Burdon T, Gibbons RJ, Garrick D & Higgs DR (2011) An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. *EMBO J*

Lyst MJ, Ekiert R, Ebert DH, Merusi C, Nowak J, Selfridge J, Guy J, Kastan NR, Robinson ND, de Lima Alves F, Rappsilber J, Greenberg ME & Bird A (2013) Rett syndrome mutations abolish the interaction of MeCP2 with the NCoR/SMRT co-repressor. *Nat Neurosci* **16:** 898–902

Macleod D, Ali R & Bird A (1998) An alternative promoter in the mouse major histocompatibility complex class II I-Abeta gene: implications for the origin of CpG islands. *Mol Cell Biol* **18:** 4433–4443

Macleod D, Charlton J, Mullins J & Bird AP (1994) Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes Dev* **8:** 2282–2292

Margueron R & Reinberg D (2011) The Polycomb complex PRC2 and its mark in life. *Nature* **469:** 343–349

Margueron R, Justin N, Ohno K, Sharpe ML, Son J, Drury WJ, Voigt P, Martin SR, Taylor WR, De Marco V, Pirrotta V, Reinberg D & Gamblin SJ (2009) Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature* **461:** 762–767

Margueron R, Li G, Sarma K, Blais A, Zavadil J, Woodcock CL, Dynlacht BD & Reinberg D (2008) Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms. *Mol Cell* **32:** 503–518

Marks H, Kalkan T, Menafra R, Denissov S, Jones K, Hofemeister H, Nichols J, Kranz A, Stewart AF, Smith A & Stunnenberg HG (2012) The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149:** 590–604

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, Turecki G, Delaney A, Varhol R, Thiessen N, Shchors K, Heine VM, Rowitch DH, Xing X, Fiore C, Schillebeeckx M, et al (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466:** 253–257

McCabe MT, Ott HM, Ganji G, Korenchuk S, Thompson C, Van Aller GS, Liu Y, Graves AP, Pietra Della A, Diaz E, LaFrance LV, Mellinger M, Duquenne C, Tian X, Kruger RG, McHugh CF, Brandt M, Miller WH, Dhanak D, Verma SK, et al (2012) EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. *Nature* **492:** 108–112

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R & Lander ES (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454:** 766–770

Mendenhall EM, Koche RP, Truong T, Zhou VW, Issac B, Chi AS, Ku M & Bernstein BE (2010) GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells. *PLoS Genet* **6:** e1001244

Meyer C, Kowarz E, Hofmann J, Renneville A, Zuna J, Trka J, Ben Abdelali R, Macintyre E, De Braekeleer E, De Braekeleer M, Delabesse E, de Oliveira MP, Cavé H, Clappier E, van Dongen JJM, Balgobind BV, van den Heuvel-Eibrink MM, Beverloo HB, Panzer-Grümayer R, Teigler-Schlegel A, et al (2009) New insights to the MLL recombinome of acute leukemias. *Leukemia* **23:** 1490–1499

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T-K, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, et al (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553–560

Miller T, Krogan NJ, Dover J, Erdjument-Bromage H, Tempst P, Johnston M, Greenblatt JF & Shilatifard A (2001) COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. *Proc Natl Acad Sci USA* **98:** 12902–12907

Milne TA, Dou Y, Martin ME, Brock HW, Roeder RG & Hess JL (2005) MLL associates specifically with a subset of transcriptionally active target genes. *Proc Natl Acad Sci U S A* **102:** 14765–14770

Milne TA, Kim J, Wang GG, Stadler SC, Basrur V, Whitcomb SJ, Wang Z, Ruthenburg AJ, Elenitoba-Johnson KSJ, Roeder RG & Allis CD (2010) Multiple interactions recruit MLL1 and MLL1 fusion proteins to the HOXA9 locus in leukemogenesis. *Mol Cell* **38:** 853–863

Mizuguchi GG, Shen XX, Landry JJ, Wu W-HW, Sen SS & Wu CC (2004) ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* **303:** 343–348

Mohan M, Herz H-M, Smith ER, Zhang Y, Jackson J, Washburn MP, Florens L, Eissenberg JC & Shilatifard A (2011) The COMPASS family of H3K4 methylases in Drosophila. *Mol Cell Biol* **31:** 4310–4318

Mohn F, Weber M, Rebhan M, Roloff T, Richter J, Stadler M, Bibel M & Schubeler D (2008) Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Mol Cell* **30:** 755–766

Montero LM, Filipski J, Gil P, Capel J, Martinez-Zapater JM & Salinas J (1992) The distribution of 5-methylcytosine in the nuclear genome of plants. *Nucleic Acids Res* **20:** 3207–3210

Morey L, Aloia L, Cozzuto L, Benitah SA & Di Croce L (2012a) RYBP and Cbx7 Define Specific Biological Functions of Polycomb Complexes in Mouse Embryonic Stem Cells. *Cell Reports*

Morey L, Pascual G, Cozzuto L, Roma G, Wutz A, Benitah SA & Di Croce L (2012b) Nonoverlapping functions of the Polycomb group Cbx family of proteins in embryonic stem cells. *Cell Stem Cell* **10:** 47–62

Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, et al (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562

Muntean AGA, Tan JJ, Sitwala KK, Huang YY, Bronstein JJ, Connelly JAJ, Basrur VV, Elenitoba-Johnson KSJK & Hess JLJ (2010) The PAF Complex Synergizes with MLL Fusion Proteins at HOX Loci to Promote Leukemogenesis. *Cancer Cell* **17:** 13–13

Müller J, Hart CM, Francis NJ, Vargas ML, Sengupta A, Wild B, Miller EL, O'Connor MB,

Kingston RE & Simon JA (2002) Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. *Cell* **111:** 197–208

Nabel CS, Jia H, Ye Y, Shen L, Goldschmidt HL, Stivers JT, Zhang Y & Kohli RM (2012) AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nat. Chem. Biol.* **8:** 751–758

Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R & Fraser P (2008) The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322:** 1717–1720

Ng HH, Robert F, Young RA & Struhl K (2003) Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Molecular Cell* **11:** 709–719

Norris D, Patel D, Kay G, Penny G, Brockdorff N, Sheardown S & Rastan S (1994) Evidence that random and imprinted Xist expression is controlled by preemptive methylation. *Cell* **77:** 41–51

O'Hagan HM, Wang W, Sen S, Destefano Shields C, Lee SS, Zhang YW, Clements EG, Cai Y, Van Neste L, Easwaran H, Casero RA, Sears CL & Baylin SB (2011) Oxidative damage targets complexes containing DNA methyltransferases, SIRT1, and polycomb members to promoter CpG Islands. *Cancer Cell* **20:** 606–619

Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, Cope L, Mohammad HP, Chen W, Daniel VC, Yu W, Berman DM, Jenuwein T, Pruitt K, Sharkis SJ, Watkins DN, Herman JG & Baylin SB (2007) A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat Genet* **39:** 237–242

Okano M, Bell D, Haber D & Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99:** 247–257

Okano M, Xie S & Li E (1998a) Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells. *Nucleic Acids Res* **26:** 2536–2540

Okano M, Xie S & Li E (1998b) Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet* **19:** 219–220

Ooi SKT, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin S-P, Allis CD, Cheng X & Bestor TH (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448:** 714–717

Palii SS, Van Emburgh BO, Sankpal UT, Brown KD & Robertson KD (2008) DNA Methylation Inhibitor 5-Aza-2'-Deoxycytidine Induces Reversible Genome-Wide DNA Damage That Is Distinctly Influenced by DNA Methyltransferases 1 and 3B. *Mol Cell Biol* **28:** 752–771

Pan G, Tian S, Nie J, Yang C, Ruotti V, Wei H, Jonsdottir GA, Stewart R & Thomson JA (2007) Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell* **1:** 299–312

Pasini D, Bracken AP, Hansen JB, Capillo M & Helin K (2007) The polycomb group protein

Suz12 is required for embryonic stem cell differentiation. *Mol Cell Biol* **27:** 3769–3779

Pasini D, Bracken AP, Jensen MR, Lazzerini Denchi E & Helin K (2004) Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. *EMBO J* **23:** 4061–4071

Pasini D, Cloos PAC, Walfridsson J, Olsson L, Bukowski J-P, Johansen JV, Bak M, Tommerup N, Rappsilber J & Helin K (2010) JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. *Nature* **464:** 306–310

Payer B & Lee JT (2008) X chromosome dosage compensation: how mammals keep the balance. *Annu Rev Genet* **42:** 733–772

Peng JC, Valouev A, Swigut T, Zhang J, Zhao Y, Sidow A & Wysocka J (2009) Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* **139:** 1290–1302

Penny GD, Kay GF, Sheardown SA, Rastan S & Brockdorff N (1996) Requirement for Xist in X chromosome inactivation. *Nature* **379:** 131–137

Plath K, Fang J, Mlynarczyk-Evans S, Cao R, Worringer K, Wang H, la Cruz de C, Otte A, Panning B & Zhang Y (2003) Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300:** 131–135

Popp C, Dean W, Feng S, Cokus SJ, Andrews S, Pellegrini M, Jacobsen SE & Reik W (2010) Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463:** 1101–1105

Proffitt JHJ, Davie JRJ, Swinton DD & Hattman SS (1984) 5-Methylcytosine is not detectable in Saccharomyces cerevisiae DNA. *Mol Cell Biol* **4:** 985–988

Qian J, Lin J & Zack DJ (2006) Characterization of binding sites of eukaryotic transcription factors. *Genomics Proteomics Bioinformatics* **4:** 67–79

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA & Wysocka J (2011) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470:** 279–283

Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C, Doty KR, Black JC, Hoffmann A, Carey M & Smale ST (2009) A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell* **138:** 114–128

Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP & Jaenisch R (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc Natl Acad Sci USA* **97:** 5237–5242

Rauch T, Wu X, Zhong X, Riggs A & Pfeifer G (2009) A human B cell methylome at 100-base pair resolution. *Proc Natl Acad Sci U S A* **106:** 671–678

Redon C, Pilch D, Rogakou E, Sedelnikova O, Newrock K & Bonner W (2002) Histone H2A variants H2AX and H2AZ. *Curr Opin Genet Dev* **12:** 162–169

Ringrose L & Paro R (2007) Polycomb/Trithorax response elements and epigenetic memory of cell identity. *Development* **134:** 223–232

Ringrose L, Lounnas V, Ehrlich L, Buchholz F, Wade R & Stewart AF (1998) Comparative kinetic analysis of FLP and cre recombinases: mathematical models for DNA binding and recombination. *J Mol Biol* **284:** 363–384

Rinn J, Kertesz M, Wang J, Squazzo S, Xu X, Brugmann S, Goodnough L, Helms J, Farnham P, Segal E & Chang H (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129:** 1311–1323

Rodriguez J, Frigola J, Vendrell E, Risques R, Fraga M, Morales C, Moreno V, Esteller M, Capella G, Ribas M & Peinado M (2006) Chromosomal instability correlates with genome-wide DNA demethylation in human primary colorectal cancers. *Cancer Res* **66:** 8462–9468

Rugg-Gunn PJ, Cox BJ, Ralston A & Rossant J (2010) Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo. *Proc Natl Acad Sci USA* **107:** 10783–10790

Rusmintratip V & Sowers LC (2000) An unexpectedly high excision capacity for mispaired 5-hydroxymethyluracil in human cell extracts. *Proc Natl Acad Sci U S A* **97:** 14183–14187

Ruthenburg AJ, Allis CD & Wysocka J (2007) Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Molecular Cell* **25:** 15–30

Ruthenburg AJ, Wang W, Graybosch DM, Li H, Allis CD, Patel DJ & Verdine GL (2006) Histone H3 recognition and presentation by the WDR5 module of the MLL1 complex. *Nat Struct Mol Biol* **13:** 704–712

Ryazanova (2012) Diverse Domains of (Cytosine-5)-DNA Methyltransferases: Structural and Functional Characterization A. Yu. Ryazanova, L. A. Abrosimova, T. S. Oretskaya and E. A. Kubareva (2012). Diverse Domains of (Cytosine-5)-DNA Methyltransferases: Structural and Functional Characterization, Methylation - From DNA, RNA and Histones to Diseases and Treatment, Prof. Anica Dricu (Ed.), ISBN: 978-953-51-0881-8, InTech, DOI: 10.5772/52046. Available from: http://www.intechopen.com/books/methylation-from-dna-rna-and-histones-to-diseases-and-treatment/diverse-domains-of-cytosine-5-dna-methyltransferases-structural-and-functional-characterization

Sachs M, Onodera C, Blaschke K, Ebata KT, Song JS & Ramalho-Santos M (2013) Bivalent Chromatin Marks Developmental Regulatory Genes in the Mouse Embryonic Germline In Vivo. *Cell Reports* **3:** 1777–1784

Saha A, Wittmeyer J & Cairns BR (2006) Chromatin remodelling: the industrial revolution of DNA around histones. *Nat Rev Mol Cell Biol* **7:** 437–447

Saito M (2002) The mCpG-binding Domain of Human MBD3 Does Not Bind to mCpG but Interacts with NuRD/Mi2 Components HDAC1 and MTA2. *Journal of Biological Chemistry* **277:** 35434–35439

Saksouk N, Avvakumov N, Champagne KS, Hung T, Doyon Y, Cayrou C, Paquet E, Ullah

M, Landry A-J, Côté V, Yang X-J, Gozani O, Kutateladze TG & Côté J (2009) HBO1 HAT complexes target chromatin throughout gene coding regions via multiple PHD finger interactions with histone H3 tail. *Mol Cell* **33:** 257–265

Santos-Rosa H, Schneider R, Bannister A, Sherriff J, Bernstein B, Emre N, Schreiber S, Mellor J & Kouzarides T (2002) Active genes are tri-methylated at K4 of histone H3. *Nature* **419:** 407–411

Sarma K, Margueron R, Ivanov A, Pirrotta V & Reinberg D (2008) Ezh2 requires PHF1 to efficiently catalyze H3 lysine 27 trimethylation in vivo. *Mol Cell Biol* **28:** 2718–2731

Sarvan S, Avdic V, Tremblay V, Chaturvedi C-P, Zhang P, Lanouette S, Blais A, Brunzelle JS, Brand M & Couture J-F (2011) Crystal structure of the trithorax group protein ASH2L reveals a forkhead-like DNA binding domain. *Nat Struct Mol Biol* **18:** 857–859

Sasai N & Defossez P-A (2009) Many paths to one goal? The proteins that recognize methylated DNA in eukaryotes. *Int. J. Dev. Biol.* **53:** 323–334

Saxonov S, Berg P & Brutlag D (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103:** 1412–1417

Schilling E & Rehli M (2007) Global, comparative analysis of tissue-specific promoter CpG methylation. *Genomics* **90:** 314–323

Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, Zimmerman J, Eden E, Yakhini Z, Ben-Shushan E, Reubinoff BE, Bergman Y, Simon I & Cedar H (2006) Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat Genet* **39:** 232–236

Schmitges FW, Prusty AB, Faty M, Stützer A, Lingaraju GM, Aiwazian J, Sack R, Hess D, Li L, Zhou S, Bunker RD, Wirth U, Bouwmeester T, Bauer A, Ly-Hartig N, Zhao K, Chan H, Gu J, Gut H, Fischle W, et al (2011) Histone methylation by PRC2 is inhibited by active chromatin marks. *Molecular Cell* **42:** 330–341

Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, Wei G & Zhao K (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132:** 887–898

Schorderet P & Duboule D (2011) Structural and functional differences in the long non-coding RNA hotair in mouse and human. *PLoS Genet* **7:** e1002071

Schuettengruber B & Cavalli G (2009) Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development* **136:** 3531–3542

Schuettengruber B, Chourrout D, Vervoort M, Leblanc B & Cavalli G (2007) Genome regulation by polycomb and trithorax proteins. *Cell* **128:** 735–745

Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, Tolhuis B, van Lohuizen M, Tanay A & Cavalli G (2009) Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. *PLoS Biol* **7:** e13

Schuettengruber BB, Martinez A-MA, Iovino NN & Cavalli GG (2011) Trithorax group

proteins: switching genes on and keeping them active. *Nat Rev Mol Cell Biol* **12:** 799–814

Schwartz S, Meshorer E & Ast G (2009) Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol* **16:** 990–995

Schwartz YB & Pirrotta V (2007) Polycomb silencing mechanisms and the management of genomic programmes. *Nat Rev Genet* **8:** 9–22

Seila AC, Core LJ, Lis JT & Sharp PA (2009) Divergent transcription: a new feature of active promoters. *Cell Cycle* **8:** 2557–2564

Seisenberger S, Andrews S, Krueger F, Arand J, Walter J, Santos F, Popp C, Thienpont B, Dean W & Reik W (2012) The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells. *Molecular Cell* **48:** 849–862

Sequeira-Mendes J, Díaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N & Gómez M (2009) Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet* **5:** e1000446

Shen X, Kim W, Fujiwara Y, Simon MD, Liu Y, Mysliwiec MR, Yuan G-C, Lee Y & Orkin SH (2009) Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells. *Cell* **139:** 1303–1314

Shen X, Liu Y, Hsu Y-J, Fujiwara Y, Kim J, Mao X, Yuan G-C & Orkin SH (2008) EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency. *Mol Cell* **32:** 491–502

Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R & Oberdoerffer S (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479:** 74–79

Silva J, Mak W, Zvetkova I, Appanah R, Nesterova T, Webster Z, Peters A, Jenuwein T, Otte A & Brockdorff N (2003) Establishment of histone h3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. *Dev Cell* **4:** 481–495

Simon JA & Kingston RE (2009) Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat Rev Mol Cell Biol* **10:** 697–708

Simon JA & Kingston RE (2013) Occupying Chromatin: Polycomb Mechanisms for Getting to Genomic Targets, Stopping Transcriptional Traffic, and Staying Put. *Molecular Cell* **49:** 808–824

Sing AA, Pannell DD, Karaiskakis AA, Sturgeon KK, Djabali MM, Ellis JJ, Lipshitz HDH & Cordes SPS (2009) A vertebrate Polycomb response element governs segmentation of the posterior hindbrain. *Cell* **138:** 885–897

Sleutels F, Zwart R & Barlow D (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415:** 810–813

Smallwood A, Estève P-O, Pradhan S & Carey M (2007) Functional cooperation between HP1 and DNMT1 mediates gene silencing. *Genes & Development* **21:** 1169–1178

Smith E & Shilatifard A (2010) The chromatin signaling pathway: diverse mechanisms of recruitment of histone-modifying enzymes and varied biological outcomes. *Molecular Cell* **40:** 689–701

Smith MM (2002) Centromeres and variant histones: what, where, when and why? *Curr Opin Cell Biol* **14:** 279–285

Smith ZD & Meissner A (2013) DNA methylation: roles in mammalian development. *Nat Rev Genet* **14:** 204–220

Smith ZD, Chan MM, Mikkelsen TS, Gu H, Gnirke A, Regev A & Meissner A (2012) A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484:** 339–344

Song J, Rechkoblit O, Bestor TH & Patel DJ (2010) Structure of DNMT1-DNA Complex Reveals a Role for Autoinhibition in Maintenance DNA Methylation. *Science*

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK & Schübeler D (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480:** 490–495

Stein R, Razin A & Cedar H (1982) In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. *Proc Natl Acad Sci U S A* **79:** 3418–3422

Steward M, Lee J, O'Donovan A, Wyatt M, Bernstein B & Shilatifard A (2006) Molecular regulation of H3K4 trimethylation by ASH2L, a shared subunit of MLL complexes. *Nat Struct Mol Biol* **13:** 852–854

Stewart AF Gene Bridges Red/ET Recombination. **:** 1–16

Strahl BD & Allis CD (2000) The language of covalent histone modifications. *Nature* **403:** 41–45

Straussman R, Nejman D, Roberts D, Steinfeld I, Blum B, Benvenisty N, Simon I, Yakhini Z & Cedar H (2009) Developmental programming of CpG island methylation profiles in the human genome. *Nat Struct Mol Biol* **16:** 564–571

Suzuki M & Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9:** 465–476

Suzuki MM, Kerr ARW, De Sousa D & Bird A (2007) CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Research* **17:** 625–631

Syeda F, Fagan RL, Wean M, Avvakumov GV, Walker JR, Xue S, Dhe-Paganon S & Brenner C (2011) The replication focus targeting sequence (RFTS) domain is a DNA-competitive inhibitor of Dnmt1. *J Biol Chem* **286:** 15344–15351

Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L & Rao A (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324:** 930–935

Takai D & Jones PA (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* **99:** 3740–3745

Takeshima H, Yamashita S, Shimazu T, Niwa T & Ushijima T (2009) The presence of RNA polymerase II, active or stalled, predicts epigenetic fate of promoter CpG islands. *Genome Research* **19:** 1974–1982

Tate CM, Lee J-H & Skalnik DG (2009) CXXC finger protein 1 contains redundant functional domains that support embryonic stem cell cytosine methylation, histone methylation, and differentiation. *Mol Cell Biol* **29:** 3817–3831

Tate CM, Lee J-H & Skalnik DG (2010) CXXC finger protein 1 restricts the Setd1A histone H3K4 methyltransferase complex to euchromatin. *FEBS J* **277:** 210–223

Tavares L, Dimitrova E, Oxley D, Webster J, Poot R, Demmers J, Bezstarosti K, Taylor S, Ura H, Koide H, Wutz A, Vidal M, Elderkin S & Brockdorff N (2012) RYBP-PRC1 Complexes Mediate H2A Ubiquitylation at Polycomb Target Sites Independently of PRC2 and H3K27me3. *Cell* **148:** 664–678

Tazi J & Bird A (1990) Alternative chromatin structure at CpG islands. *Cell* **60:** 909–920

Terzi N, Churchman LS, Vasiljeva L, Weissman J & Buratowski S (2011) H3K4 trimethylation by Set1 promotes efficient termination by the Nrd1-Nab3-Sen1 pathway. *Mol Cell Biol* **31:** 3569–3583

Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr ARW, Deaton A, Andrews R, James KD, Turner DJ, Illingworth R & Bird A (2010) CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464:** 1082–1086

Tkachuk DC, Kohler S & Cleary ML (1992) Involvement of a homolog of Drosophila trithorax by 11q23 chromosomal translocations in acute leukemias. *Cell* **71:** 691–700

Tonna S, El-Osta A, Cooper ME & Tikellis C (2010) Metabolic memory and diabetic nephropathy: potential role for epigenetic mechanisms. *Nat Rev Nephrol* **6:** 332–341

Tsai M-C, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E & Chang HY (2010) Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. *Science* **329:** 689–693

Tsukada Y-I, Fang J, Erdjument-Bromage H, Warren ME, Borchers CH, Tempst P & Zhang Y (2006) Histone demethylation by a family of JmjC domain-containing proteins. *Nature* **439:** 811–816

Tsumura A, Hayakawa T, Kumaki Y, Takebayashi S-I, Sakaue M, Matsuoka C, Shimotohno K, Ishikawa F, Li E, Ueda HR, Nakayama J-I & Okano M (2006) Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. *Genes Cells* **11:** 805–814

Tweedie S, Charlton J, Clark V & Bird A (1997) Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Mol Cell Biol* **17:** 1469–1475

Valinluck V & Sowers LC (2007) Endogenous cytosine damage products alter the site

selectivity of human DNA maintenance methyltransferase DNMT1. *Cancer Res* **67:** 946–950

van Dijk EL, Chen CL, d'Aubenton-Carafa Y, Gourvennec S, Kwapisz M, Roche V, Bertrand C, Silvain M, Legoix-Né P, Loeillet S, Nicolas A, Thermes C & Morillon A (2011) XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature* **475:** 114–117

Vardimon L, Kressmann A, Cedar H, Maechler M & Doerfler W (1982) Expression of a cloned adenovirus gene is inhibited by in vitro methylation. *Proc Natl Acad Sci U S A* **79:** 1073–1077

Vastenhouw NL & Schier AF (2012) Bivalent histone modifications in early embryogenesis. *Curr Opin Cell Biol* **24:** 374–386

Vastenhouw NL, Zhang Y, Woods IG, Imam F, Regev A, Liu XS, Rinn J & Schier AF (2010) Chromatin signature of embryonic pluripotency is established during genome activation. *Nature* **464:** 922–926

Vermeulen M, Mulder KW, Denissov S, Pijnappel WWMP, van Schaik FMA, Varier RA, Baltissen MPA, Stunnenberg HG, Mann M & Timmers HTM (2007) Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131:** 58–69

Viré E, Brenner C, Deplus R, Blanchon L, Fraga M, Didelot C, Morey L, Van Eynde A, Bernard D, Vanderwinden J-M, Bollen M, Esteller M, Di Croce L, de Launoit Y & Fuks F (2006) The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439:** 871–874

Voigt P, Leroy G, Drury WJ, Zee BM, Son J, Beck DB, Young NL, Garcia BA & Reinberg D (2012) Asymmetrically modified nucleosomes. *Cell* **151:** 181–193

Voigt P, Tee WW & Reinberg D (2013) A double take on bivalent promoters. *Genes & Development* **27:** 1318–1338

Voo KS, Carlone DL, Jacobsen BM, Flodin A & Skalnik DG (2000) Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Mol Cell Biol* **20:** 2108–2121

Walker E, Chang WY, Hunkapiller J, Cagney G, Garcha K, Torchia J, Krogan NJ, Reiter JF & Stanford WL (2010) Polycomb-like 2 associates with PRC2 and regulates transcriptional networks during mouse embryonic stem cell self-renewal and differentiation. *Cell Stem Cell* **6:** 153–166

Walsh CP, Chaillet JR & Bestor TH (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* **20:** 116–117

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, Wysocka J, Lei M, Dekker J, Helms JA & Chang HY (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472:** 120–124

Wang P, Lin C, Smith ER, Guo H, Sanderson BW, Wu M, Gogol M, Alexander T, Seidel C, Wiedemann LM, Ge K, Krumlauf R & Shilatifard A (2009) Global analysis of H3K4 methylation defines MLL family member targets and points to a role for MLL1-mediated H3K4 methylation in the regulation of transcriptional initiation by RNA polymerase II. *Mol Cell Biol* **29:** 6074–6085

Weber M, Hellmann I, Stadler M, Ramos L, Paabo S, Rebhan M & Schubeler D (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* **39:** 457–466

Williams K, Christensen J, Pedersen MT, Johansen JV, Cloos PAC, Rappsilber J & Helin K (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473:** 343–348

Wilson AS, Power BE & Molloy PL (2007) DNA hypomethylation and human diseases. *Biochim Biophys Acta* **1775:** 138–162

Wood A, Schneider J, Dover J, Johnston M & Shilatifard A (2003) The Paf1 complex is essential for histone monoubiquitination by the Rad6-Bre1 complex, which signals for histone methylation by COMPASS and Dot1p. *J Biol Chem* **278:** 34739–34742

Woodcock DM, Lawler CB, Linsenmeyer ME, Doherty JP & Warren WD (1997) Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. *J Biol Chem* **272:** 7810–7816

Wossidlo M, Nakamura T, Lepikhov K, Marques CJ, Zakhartchenko V, Boiani M, Arand J, Nakano T, Reik W & Walter J (2011) 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat Comms* **2:** 241

Wu H, Coskun V, Tao J, Xie W, Ge W, Yoshikawa K, Li E, Zhang Y & Sun YE (2010a) Dnmt3a-Dependent Nonpromoter DNA Methylation Facilitates Transcription of Neurogenic Genes. *Science* **329:** 444–448

Wu H, D'Alessio AC, Ito S, Wang Z, Cui K, Zhao K, Sun YE & Zhang Y (2011a) Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes & Development* **25:** 679–684

Wu H, D'Alessio AC, Ito S, Xia K, Wang Z, Cui K, Zhao K, Sun YE & Zhang Y (2011b) Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature*

Wu HH, Coskun VV, Tao JJ, Xie WW, Ge WW, Yoshikawa KK, Li EE, Zhang YY & Sun YEY (2010b) Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. *Science* **329:** 444–448

Wu M, Wang PF, Lee JS, Martin-Brown S, Florens L, Washburn M & Shilatifard A (2008) Molecular regulation of H3K4 trimethylation by Wdr82, a component of human Set1/COMPASS. *Mol Cell Biol* **28:** 7337–7344

Wu X, Johansen JV & Helin K (2013) Fbxl10/Kdm2b Recruits Polycomb Repressive Complex 1 to CpG Islands and Regulates H2A Ubiquitylation. *Molecular Cell* **49:** 1134–1146

Wutz A, Smrzka O, Schweifer N, Schellander K, Wagner E & Barlow D (1997) Imprinted expression of the Igf2r gene depends on an intronic CpG island. *Nature* **389:** 745–749

Wysocka J, Swigut T, Xiao H, Milne TA, Kwon SY, Landry J, Kauer M, Tackett AJ, Chait BT, Badenhorst P, Wu C & Allis CD (2006) A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* **442:** 86–90

Xi HH, Yu YY, Fu YY, Foley JJ, Halees AA & Weng ZZ (2007) Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genes & Development* **17:** 798–806

Xu C, Bian C, Lam R, Dong A & Min J (2011a) The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain. *Nat Comms* **2:** 227

Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, Hulten M, Qu X, Russo JJ & Viegas-Péquignot E (1999) Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **402:** 187–191

Xu Y, Wu F, Tan L, Kong L, Xiong L, Deng J, Barbera AJ, Zheng L, Zhang H, Huang S, Min J, Nicholson T, Chen T, Xu G, Shi Y, Zhang K & Shi YG (2011b) Genome-wide Regulation of 5hmC, 5mC, and Gene Expression by Tet1 Hydroxylase in Mouse Embryonic Stem Cells. *Molecular Cell* **42:** 451–464

Xu Y, Xu C, Kato A, Tempel W, Abreu JG, Bian C, Hu Y, Hu D, Zhao B, Cerovina T, Diao J, Wu F, He HH, Cui Q, Clark E, Ma C, Barbara A, Veenstra GJC, Xu G, Kaiser UB, et al (2012) Tet3 CXXC Domain and Dioxygenase Activity Cooperatively Regulate Key Genes for Xenopus Eye and Neural Development. *Cell* **151:** 1200–1213

Yang ASA, Doshi KDK, Choi S-WS, Mason JBJ, Mannari RKR, Gharybian VV, Luna RR, Rashid AA, Shen LL, Estecio MRHM, Kantarjian HMH, Garcia-Manero GG & Issa J-PJJ (2006) DNA methylation changes after 5-aza-2'-deoxycytidine therapy in patients with leukemia. *Cancer Res* **66:** 5495–5503

Yang Y & Hua X (2007) In search of tumor suppressing functions of menin. *Mol Cell Endocrinol* **265-266:** 34–41

Yap KL, Li S, Muñoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, Gil J, Walsh MJ & Zhou M-M (2010) Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol Cell* **38:** 662–674

Yoder JA, Soman NS, Verdine GL & Bestor TH (1997) DNA (cytosine-5)-methyltransferases in mouse cells and tissues. Studies with a mechanism-based probe. *J Mol Biol* **270:** 385–395

Young NL, DiMaggio PA, Plazas-Mayorca MD, Baliban RC, Floudas CA & Garcia BA (2009) High throughput characterization of combinatorial histone codes. *Mol. Cell Proteomics* **8:** 2266–2284

Yu BD, Hess JL, Horning SE, Brown GA & Korsmeyer SJ (1995) Altered Hox expression and segmental identity in Mll-mutant mice. *Nature* **378:** 505–508

Yuan WW, Wu TT, Fu HH, Dai CC, Wu HH, Liu NN, Li XX, Xu MM, Zhang ZZ, Niu TT,

Han ZZ, Chai JJ, Zhou XJX, Gao SS & Zhu BB (2012) Dense chromatin activates Polycomb repressive complex 2 to regulate H3 lysine 27 methylation. *Science* **337:** 971–975

Zhang H, Zhang X, Clark E, Mulcahey M, Huang S & Shi YG (2010a) TET1 is a DNA-binding protein that modulates DNA methylation and gene transcription via hydroxylation of 5-methylcytosine. *Cell Res* **20:** 1390–1393

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SWL, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE & Ecker JR (2006) Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis. *Cell* **126:** 1189–1201

Zhang Y, Buchholz F, Muyrers JP & Stewart AF (1998) A new logic for DNA engineering using recombination in Escherichia coli. *Nat Genet* **20:** 123–128

Zhang Y, Jurkowska R, Soeroes S, Rajavelu A, Dhayalan A, Bock I, Rathert P, Brandt O, Reinhardt R, Fischle W & Jeltsch A (2010b) Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. *Nucleic Acids Res* **38:** 4246–4253

Zhang Y, Ng HH, Erdjument-Bromage H, Tempst P, Bird A & Reinberg D (1999) Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes & Development* **13:** 1924–1935

Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M & Lee JT (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular Cell* **40:** 939–953

Zhao J, Sun BK, Erwin JA, Song JJ & Lee JT (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322:** 750–756

Zhao XD, Han X, Chew JL, Liu J, Chiu KP, Choo A, Orlov YL, Sung W-K, Shahab A, Kuznetsov VA, Bourque G, Oh S, Ruan Y, Ng HH & Wei C-L (2007) Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* **1:** 286–298

Zhou JC, Blackledge NP, Farcas AM & Klose RJ (2011) Recognition of CpG island chromatin by KDM2A requires direct and specific interaction with linker DNA. *Mol Cell Biol*

Zhu J, He F, Hu S & Yu J (2008) On the nature of human housekeeping genes. *Trends Genet* **24:** 481–484

Zilberman D, Coleman-Derr D, Ballinger T & Henikoff S (2008) Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature* **456:** 125–129